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

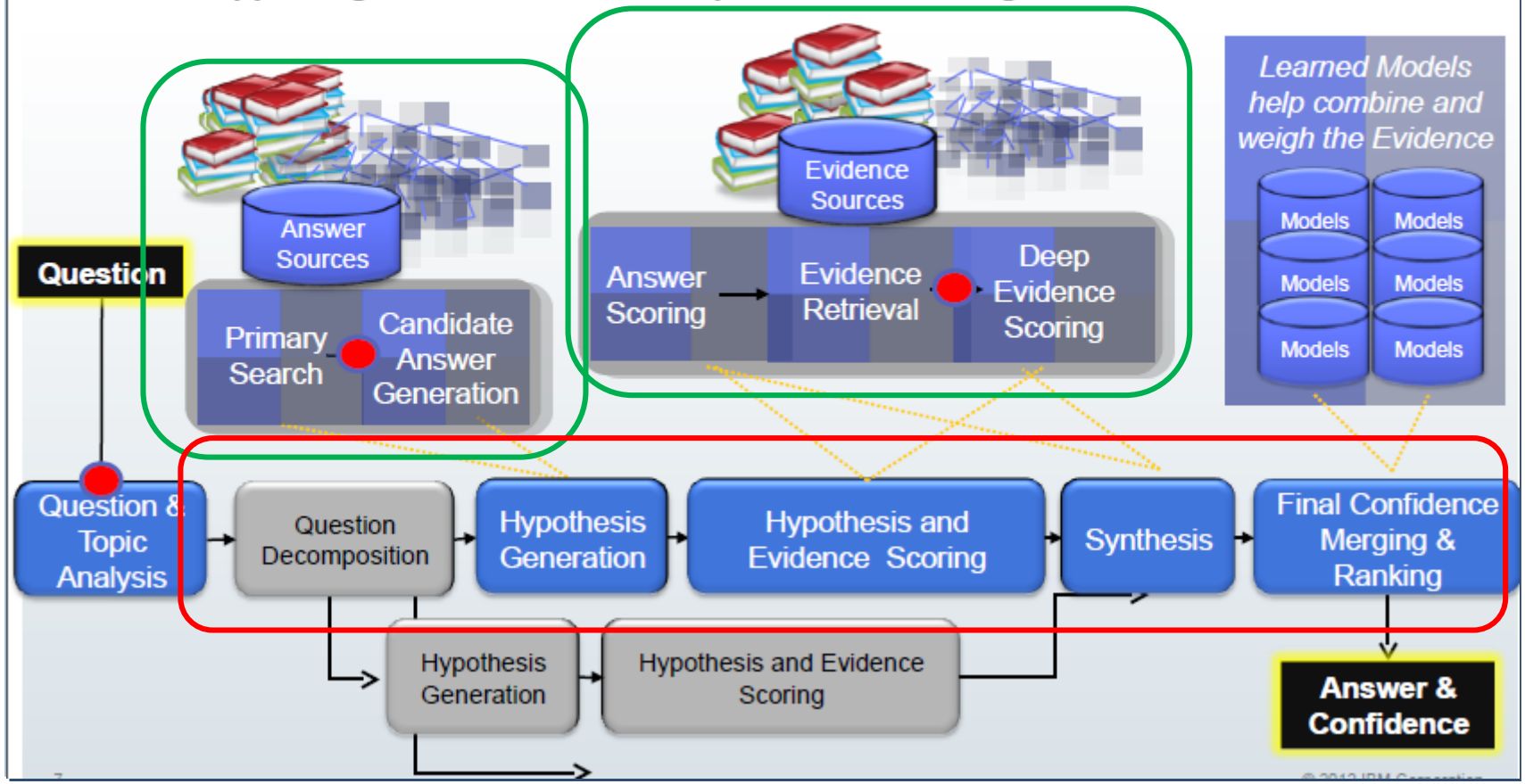
- Goal-oriented (effectiveness driven)
  - Any process that generates useful results that are non-obvious is called “mining”.
  - Keywords: “**useful**” + “**non-obvious**”
  - Data isn’t necessarily massive
- Method-oriented (efficiency driven)
  - Any process that involves extracting information from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

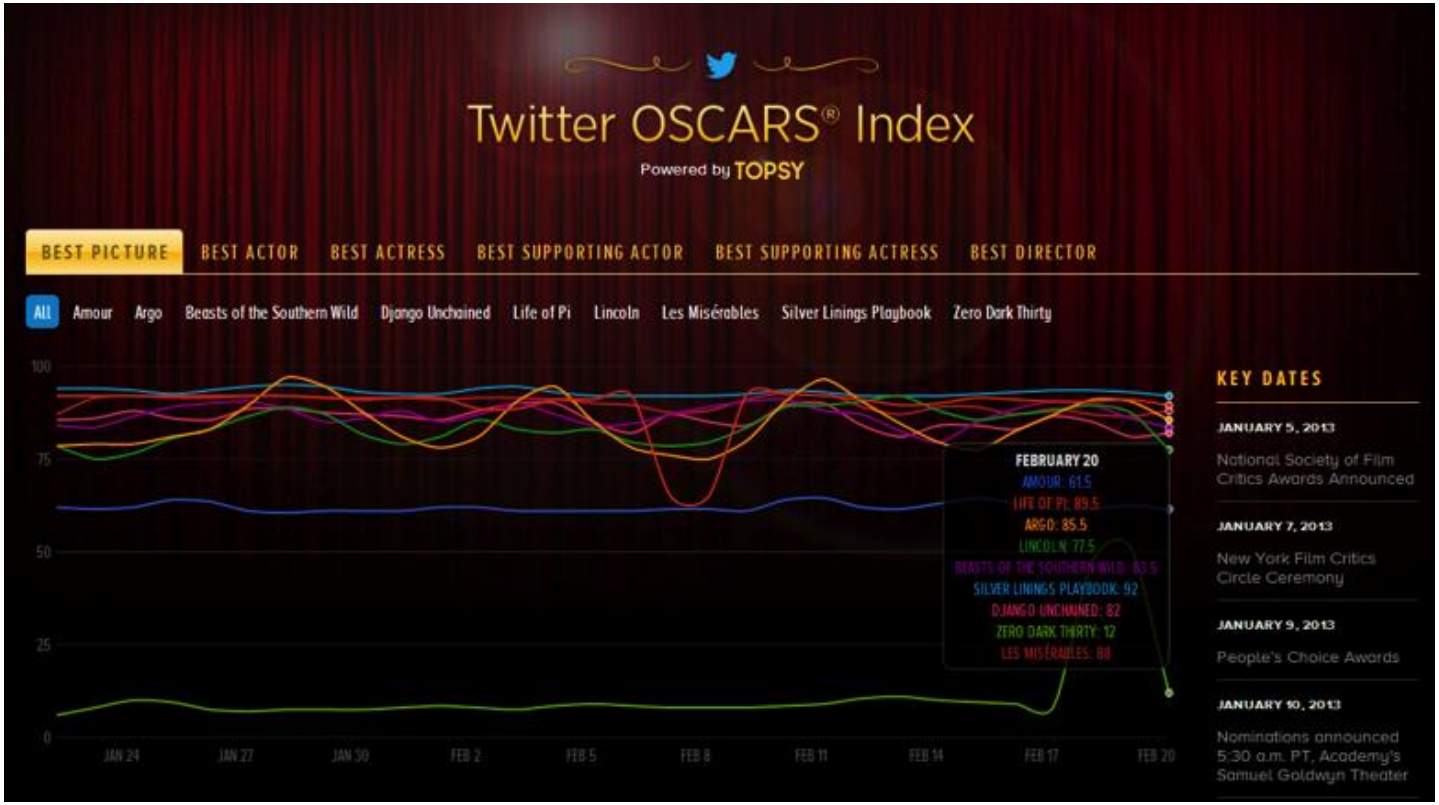
- *"Watson had access to <u>200 million pages</u> of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia" – PC World*
- *"The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary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

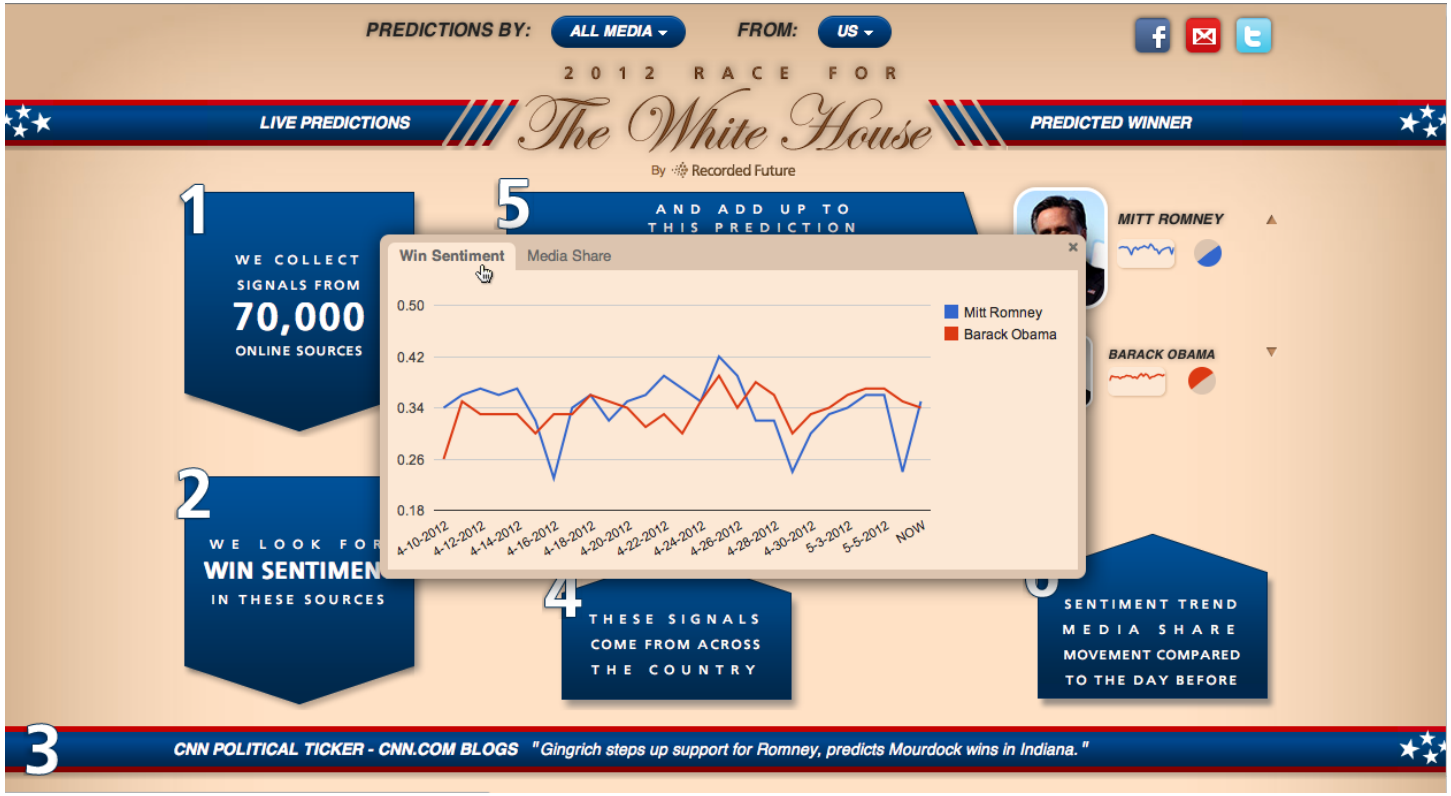
Text mining around us

- Sentiment analysis



Text mining around us

- Sentiment analysis

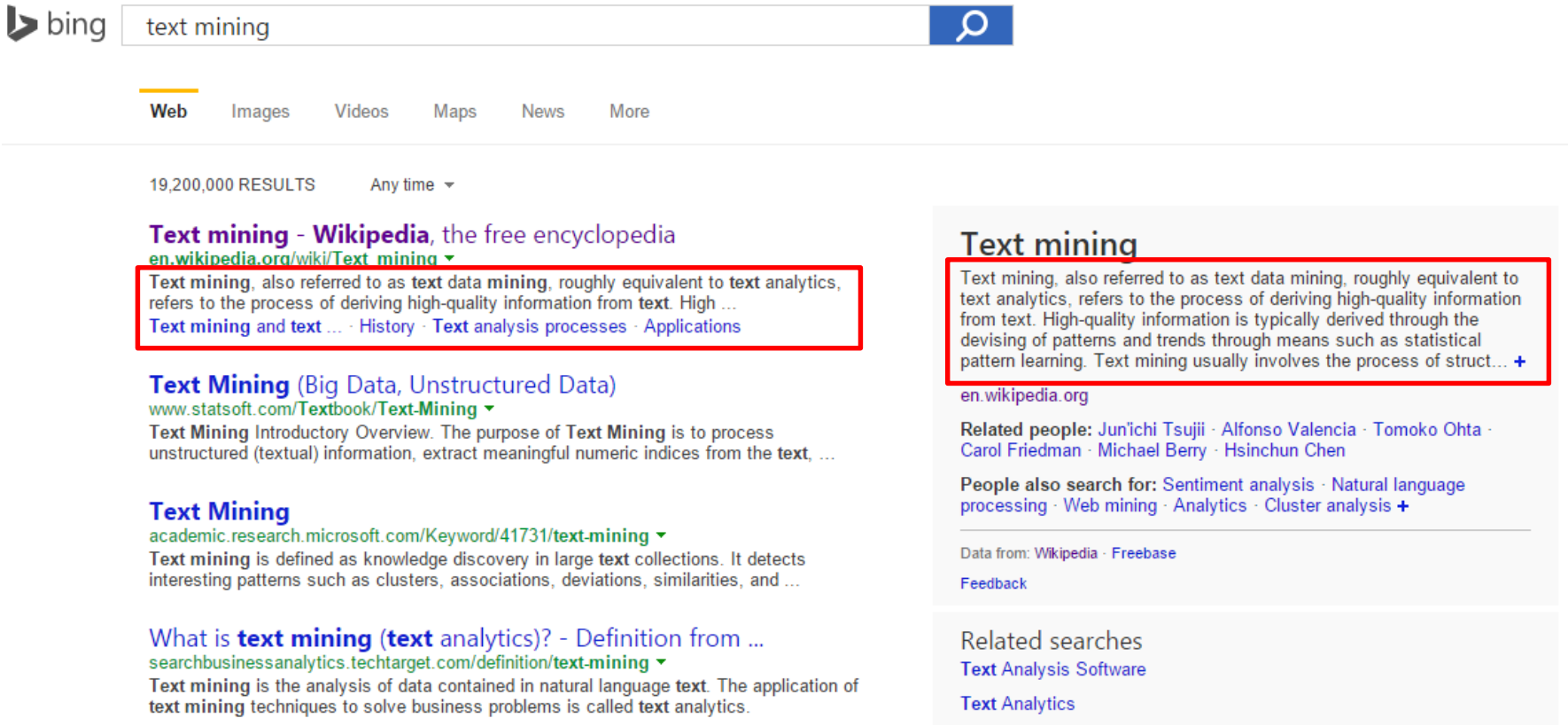


Text mining around us

- Document summarization

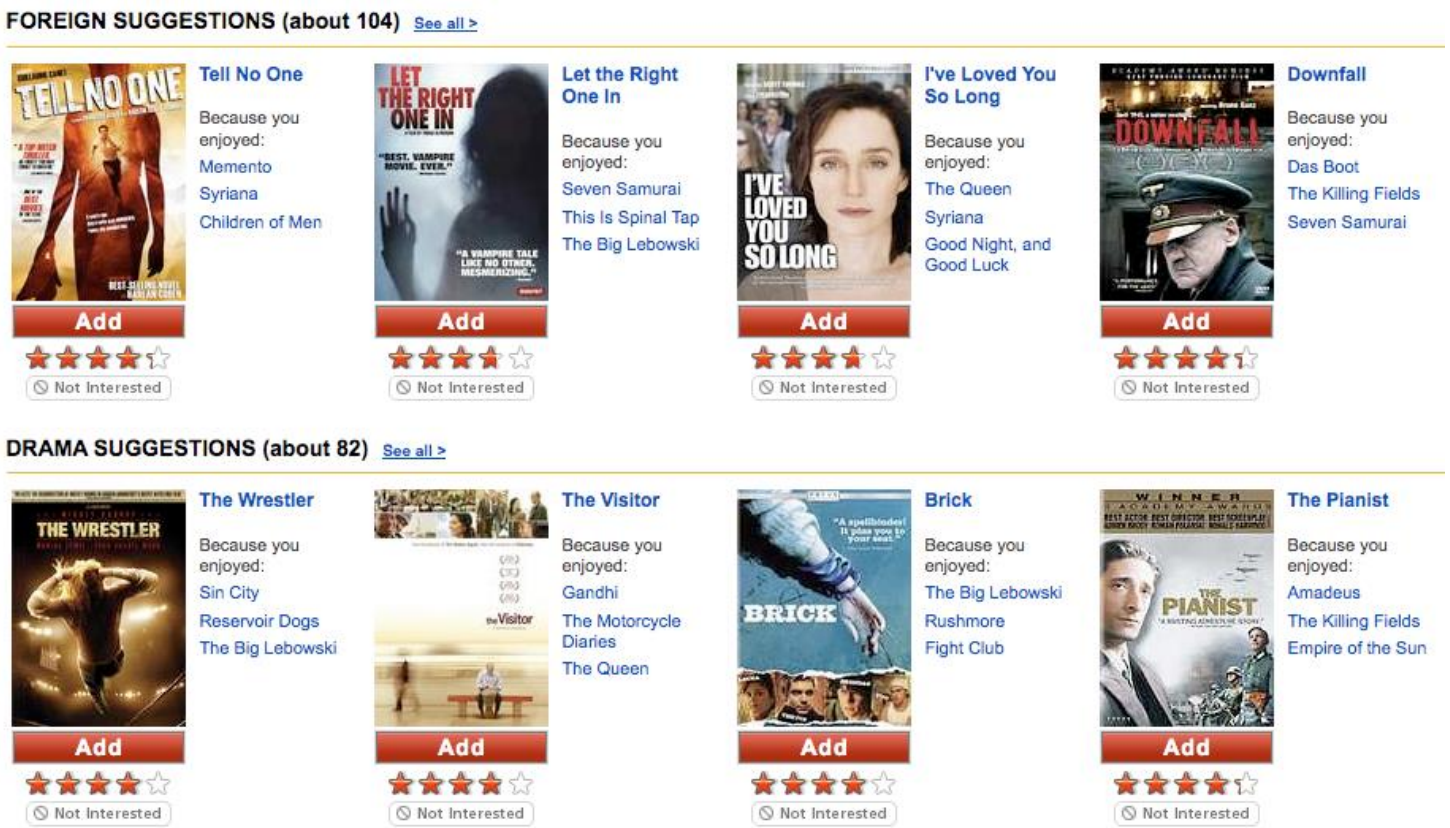
Text mining around us

- Document summarization



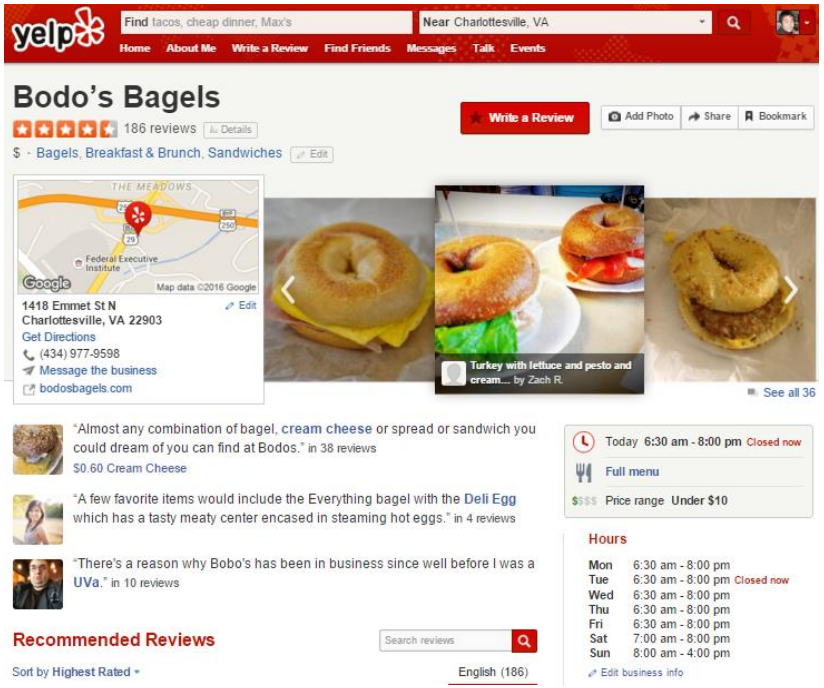
Text mining around us

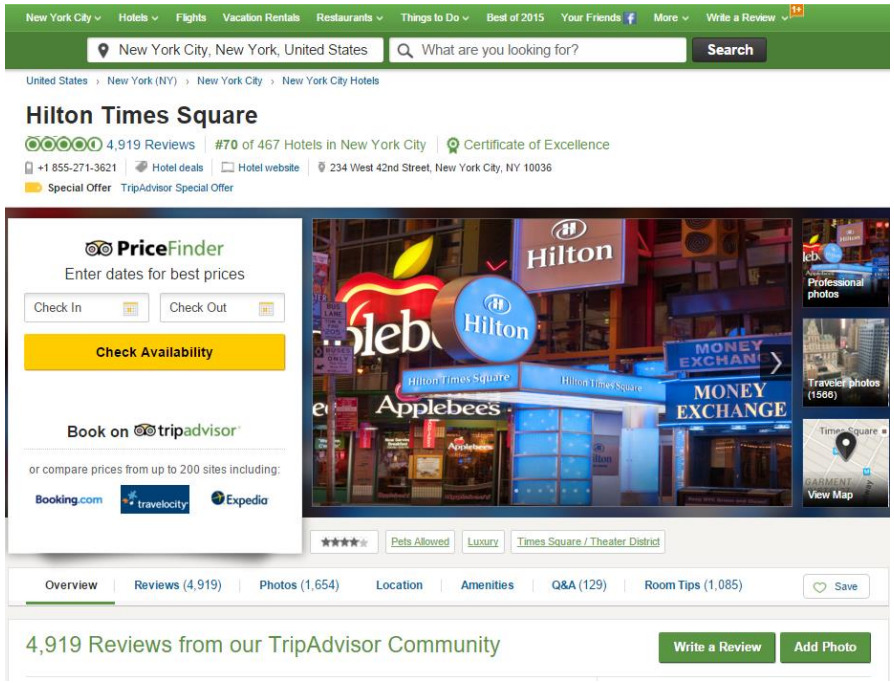
- Movie recommendation



Text mining around us

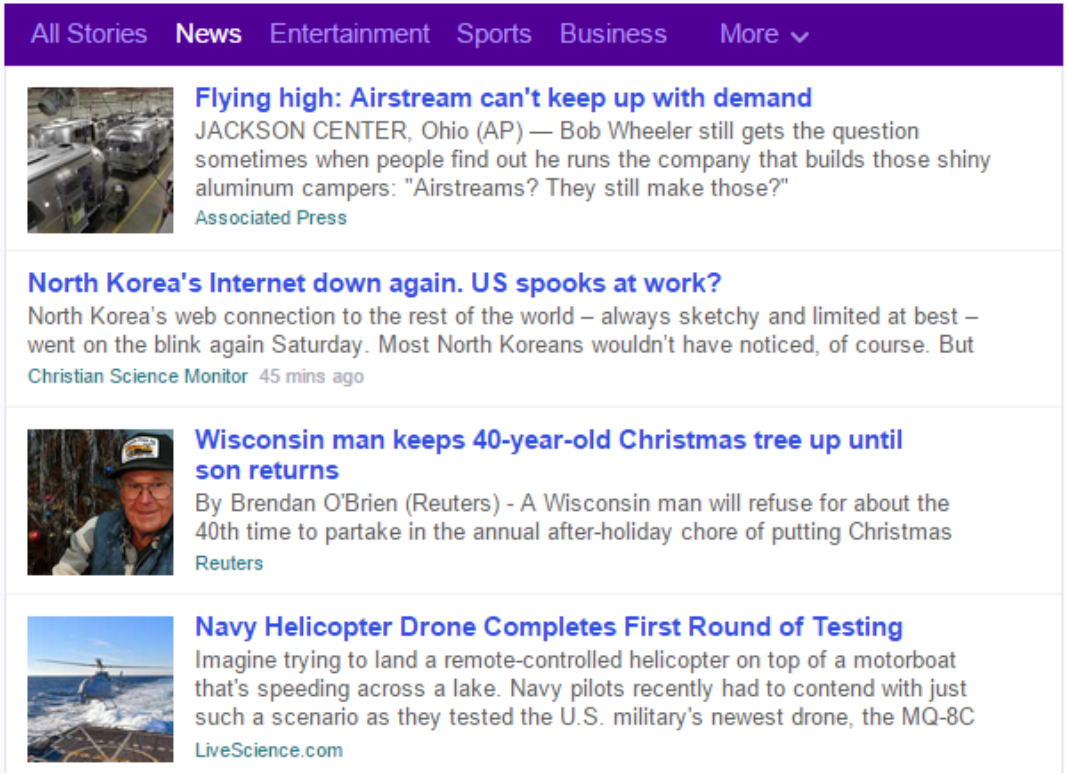
- Restaurant/hotel recommendation





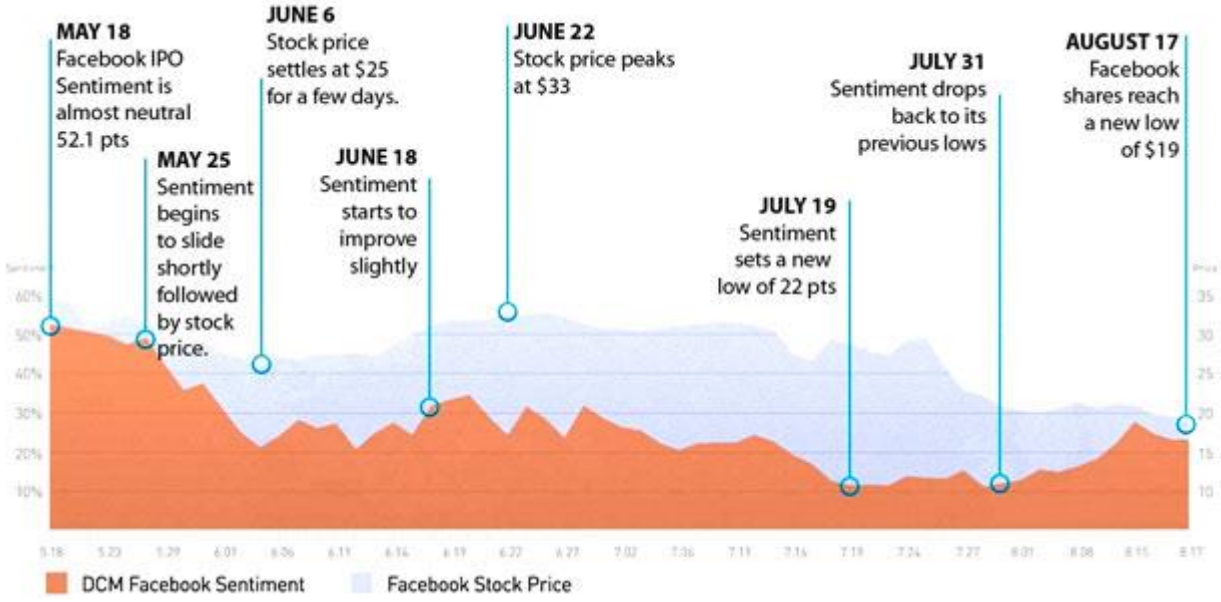
Text mining around us

- News recommendation



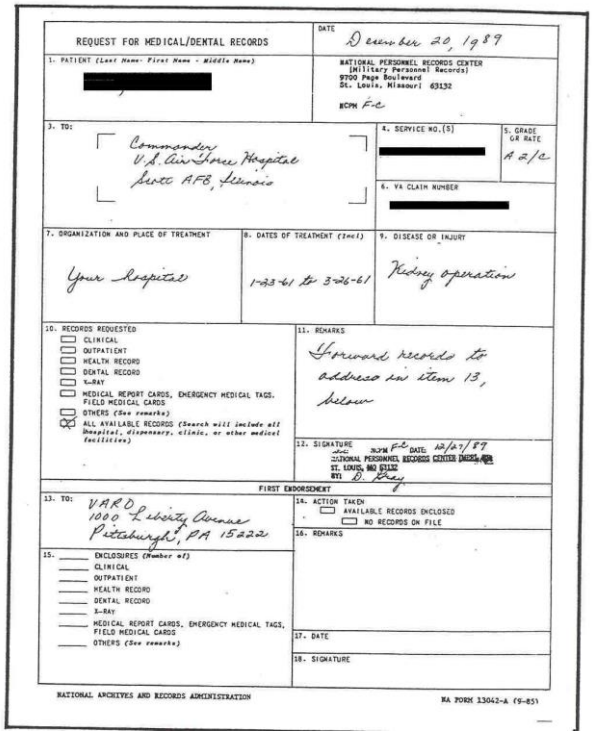
Text mining around us

- Text analytics in financial services



Text mining around us

- Text analytics in healthcare



REQUEST FOR MEDICAL/DENTAL RECORDS

DATE December 20, 1989

1. PATIENT (Last Name - First Name - Middle Name) [redacted]

NATIONAL PERSONNEL RECORDS CENTER (Military Personnel Records) 9700 Page Boulevard St. Louis, Missouri 63132

NCPM F-C

3. TO: Commander V.A. Air Force Hospital Scott AFB, Illinois

4. SERVICE NO. (S) [redacted]

5. GRADE OR RATE A 2/c

6. VA CLAIM NUMBER [redacted]

7. ORGANIZATION AND PLACE OF TREATMENT: Your Hospital

8. DATES OF TREATMENT (Incl) 1-23-61 to 3-26-61

9. DISEASE OR INJURY: Kidney operation

10. RECORDS REQUESTED
- [ ] CLINICAL
- [ ] OUTPATIENT
- [ ] HEALTH RECORD
- [ ] DENTAL RECORD
- [ ] X-RAY
- [ ] MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
- [ ] OTHERS (See remarks)
- [x] ALL AVAILABLE RECORDS (Search will include all hospital, dispensary, clinic, or other medical facilities)

11. REMARKS: Forward records to address in item 13, below

12. SIGNATURE NCPM F-C DATE 12/27/89 NATIONAL PERSONNEL RECORDS CENTER (MPR) ST. LOUIS, MO 63132 BY: D. Gray

FIRST ENDORSEMENT

13. TO: VARO 1000 Liberty Avenue Pittsburgh, PA 15222

14. ACTION TAKEN
- [ ] AVAILABLE RECORDS ENCLOSED
- [ ] NO RECORDS ON FILE

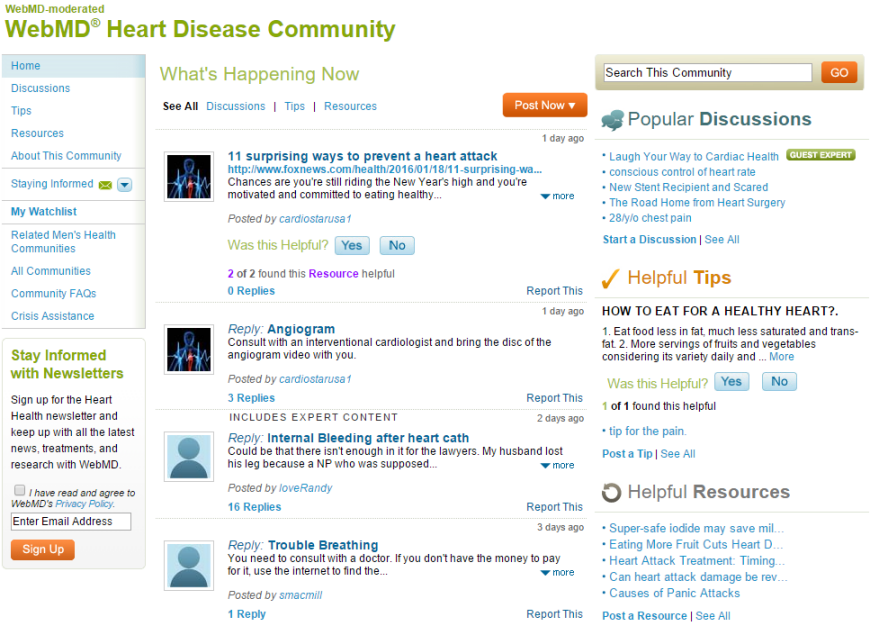
15. ____ ENCLOSURES (Number of)
- ____ CLINICAL
- ____ OUTPATIENT
- ____ HEALTH RECORD
- ____ DENTAL RECORD
- ____ X-RAY
- ____ MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
- ____ OTHERS (See remarks)

16. REMARKS

17. DATE

18. SIGNATURE

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION — NA 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

1 day ago
11 surprising ways to prevent a heart attack
http://www.foxnews.com/health/2016/01/18/11-surprising-wa...
Chances are you're still riding the New Year's high and you're motivated and committed to eating healthy... more
Posted by cardiostarusa1
Was this Helpful? Yes No
2 of 2 found this Resource helpful
0 Replies Report This

1 day ago
Reply: Angiogram
Consult with an interventional cardiologist and bring the disc of the angiogram video with you.
Posted by cardiostarusa1
3 Replies Report This

INCLUDES EXPERT CONTENT 2 days ago
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... more
Posted by loveRandy
16 Replies Report This

3 days ago
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... more
Posted by smacmill
1 Reply Report This

Search This Community GO

Popular Discussions
- Laugh Your Way to Cardiac Health GUEST EXPERT
- conscious control of heart rate
- New Stent Recipient and Scared
- The Road Home from Heart Surgery
- 28/y/o chest pain

Start a Discussion | See All

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? Yes No
1 of 1 found this helpful
- tip for the pain.

Post a Tip | See All

Helpful Resources
- Super-safe iodide 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

Applied machine learning
Natural language processing
Information retrieval
Emails
Software documentations
Blogs
Web pages
Tweets
Scientific literature
News articles

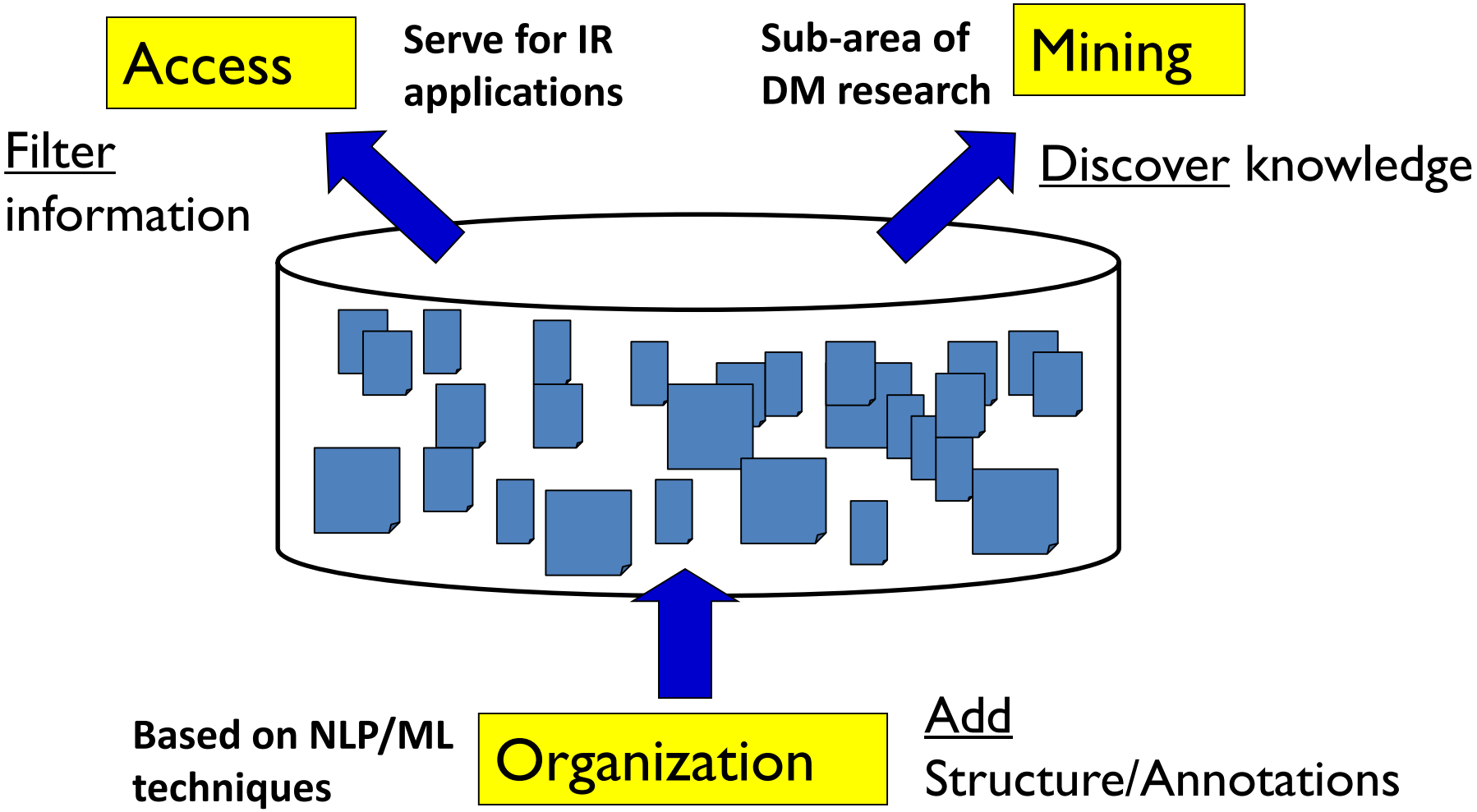
Text mining v.s. NLP, IR, DM…

- 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

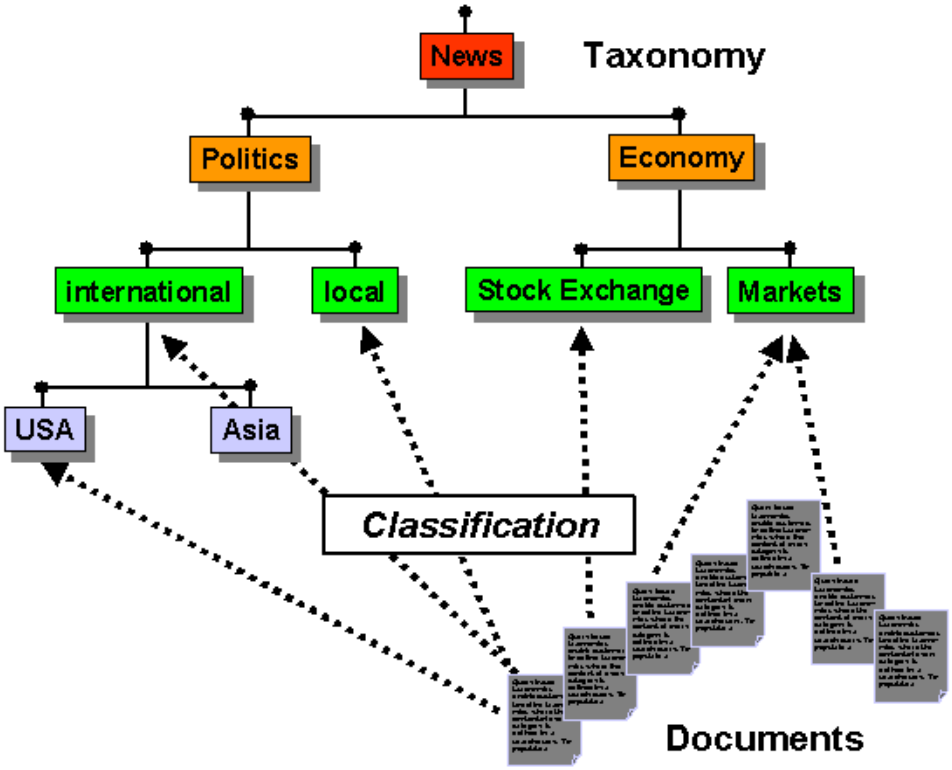


Challenges in text mining

- 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?

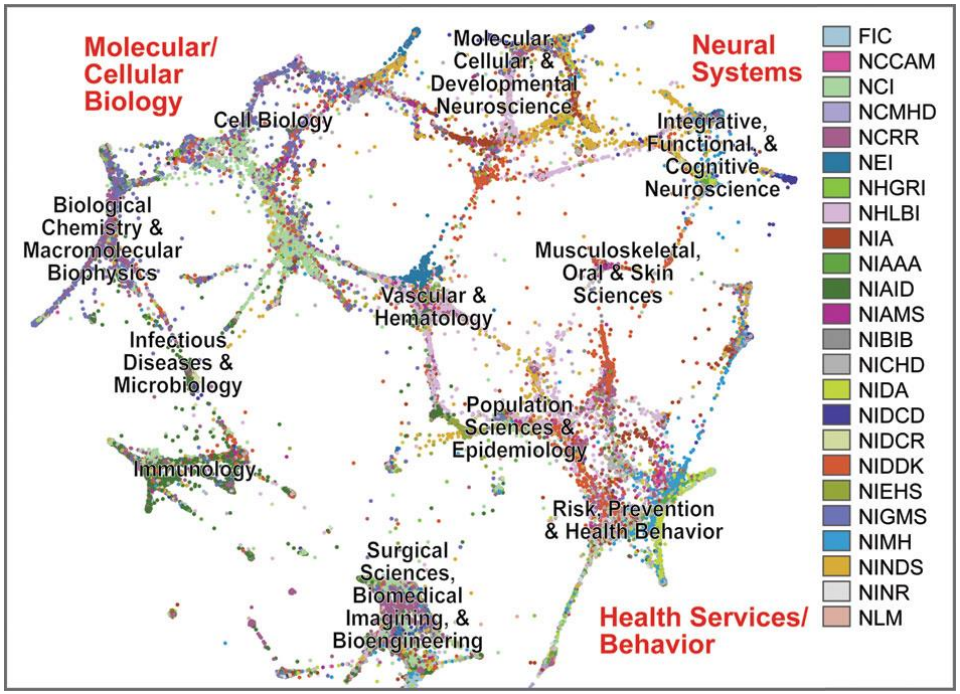
Text mining problems we will solve

- Document categorization
  - Adding structures to the text corpus



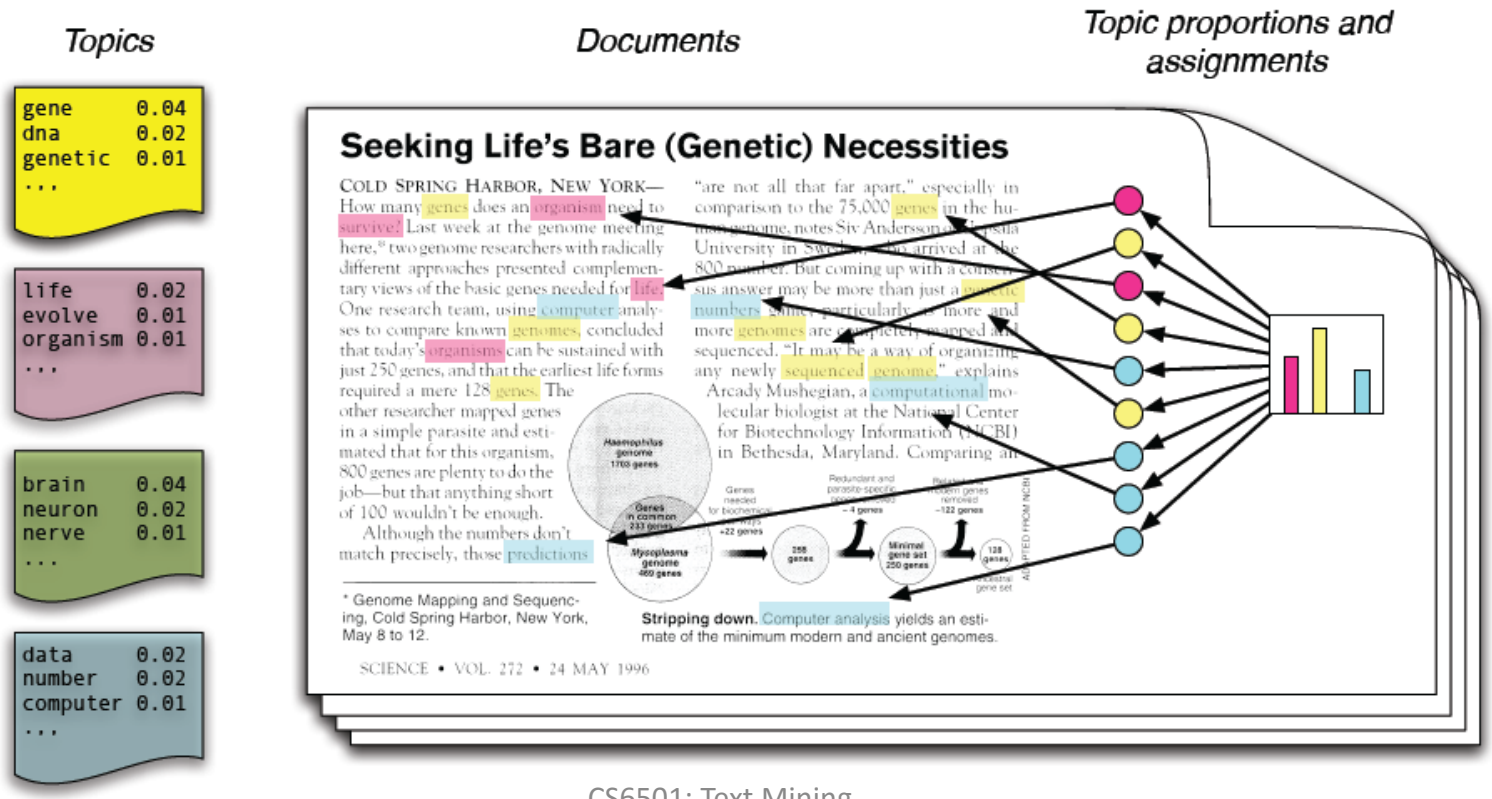
Text mining problems we will solve

- Text clustering
  - Identifying structures in the text corpus



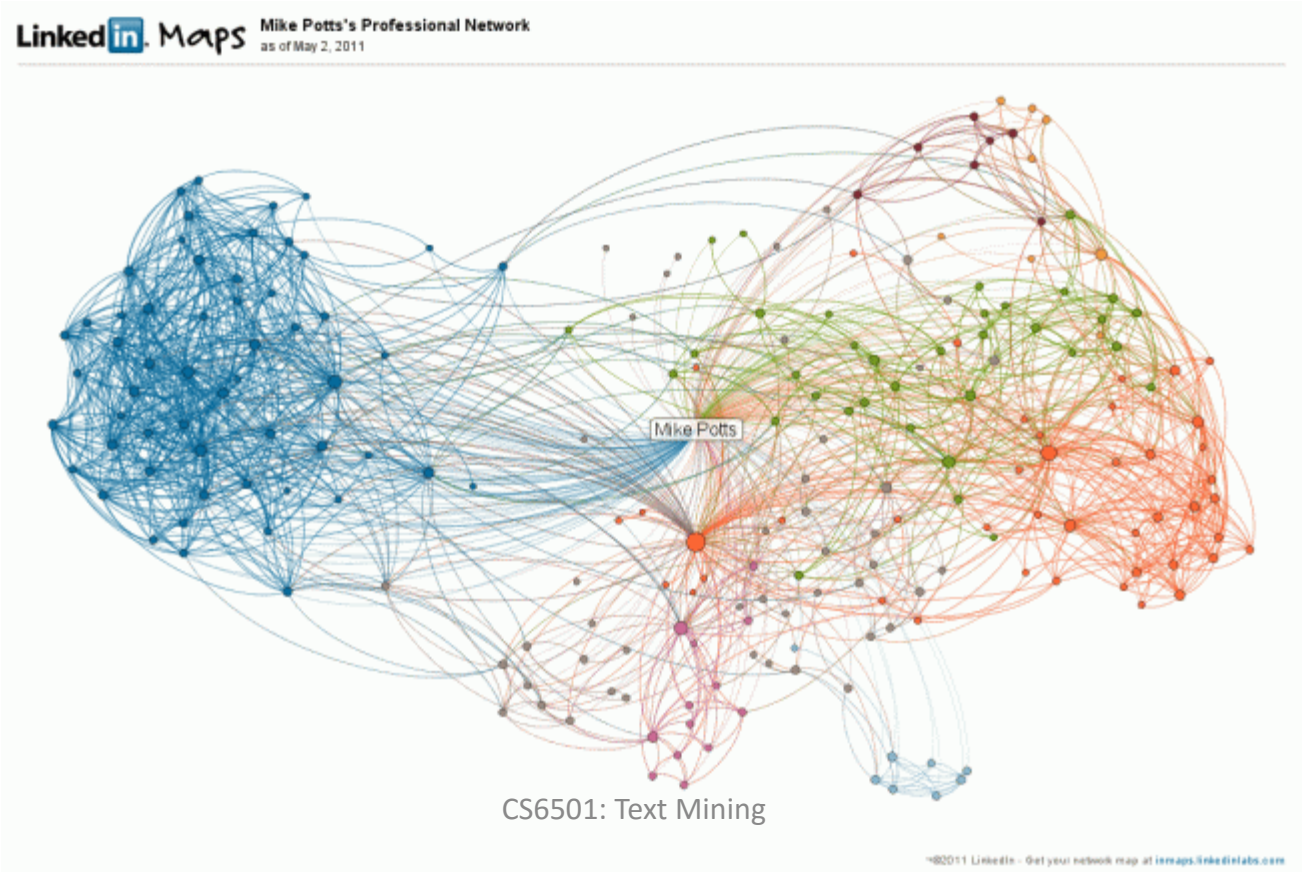
Text mining problems we will solve

- Topic modeling
  - Identifying structures in the text corpus

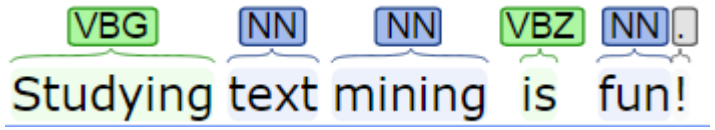


Text mining problems we will solve

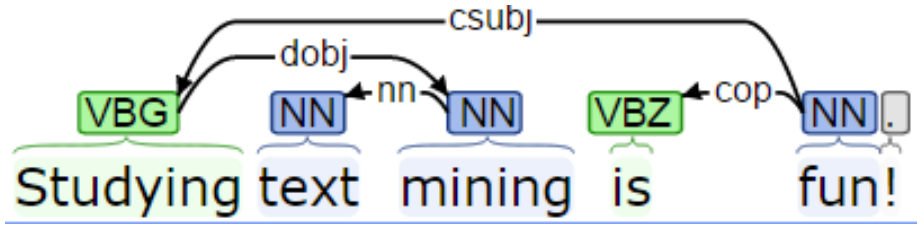
- Social media and network analysis
  - Exploring additional structure in the text corpus



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

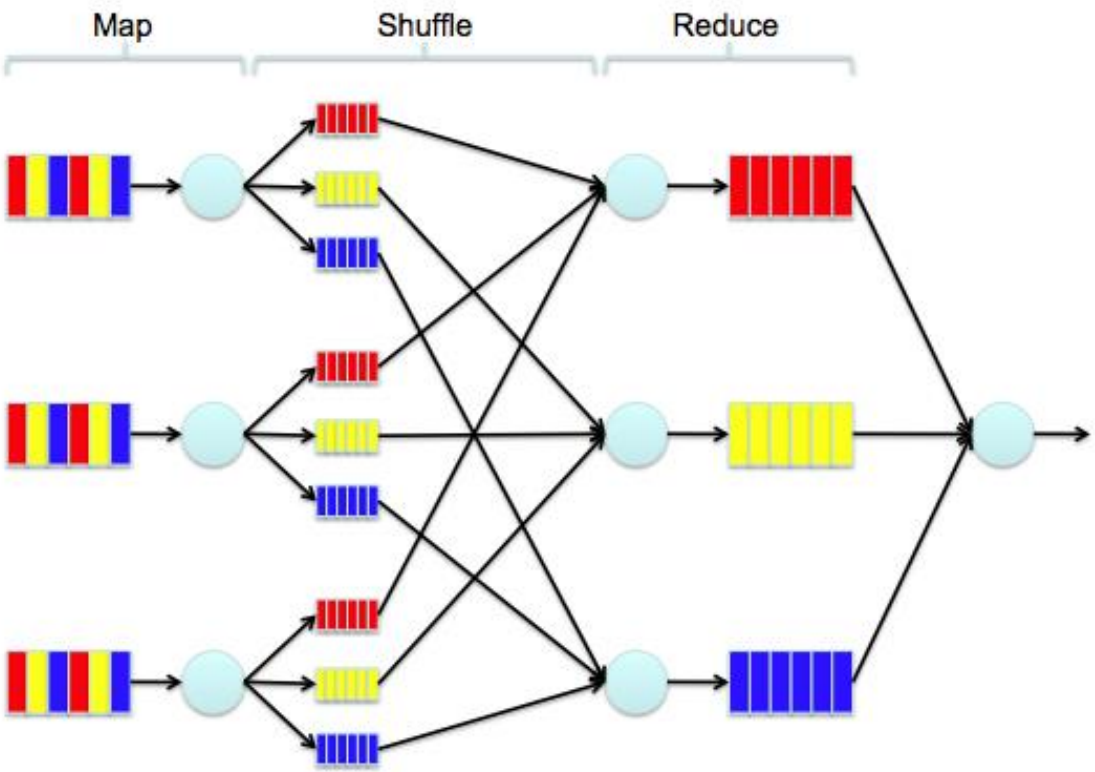
- Machine learning techniques
  - Supervised methods
    - Naïve Bayes, k Nearest Neighbors, Logistic Regression
  - Unsupervised methods
    - K-Means, hierarchical clustering, topic models
  - Semi-supervised methods
    - Expectation Maximization

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

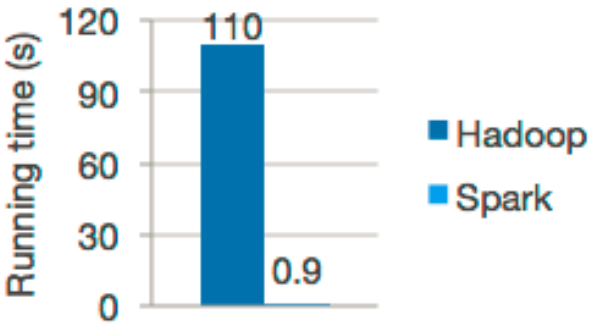
Scalability is crucial

- Large scale text processing techniques
  - MapReduce framework



State-of-the-art solutions

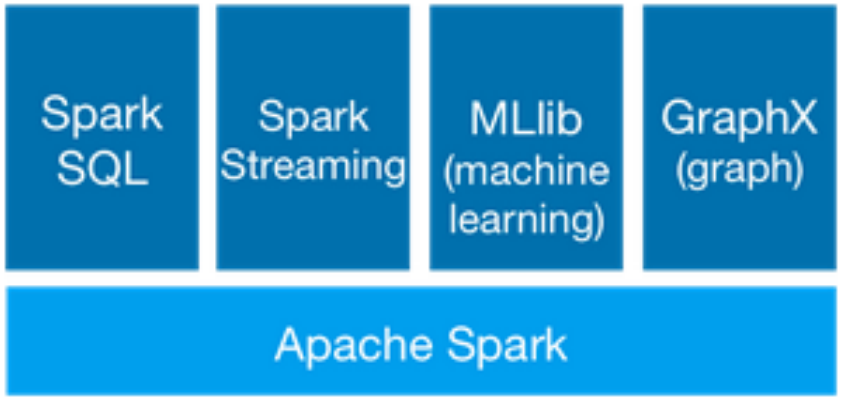
- Apache Spark (spark.apache.org)
  - In-memory MapReduce
    - Specialized for machine learning algorithms
  - Speed
    - 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.



Logistic regression in Hadoop and Spark

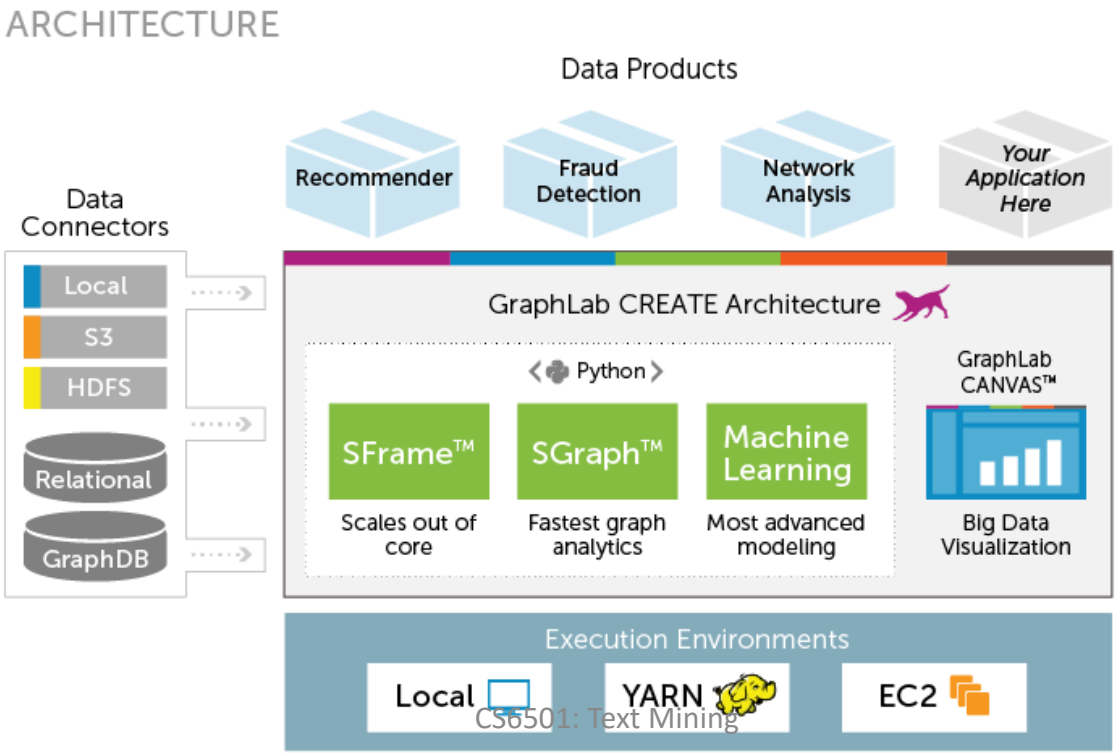
State-of-the-art solutions

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



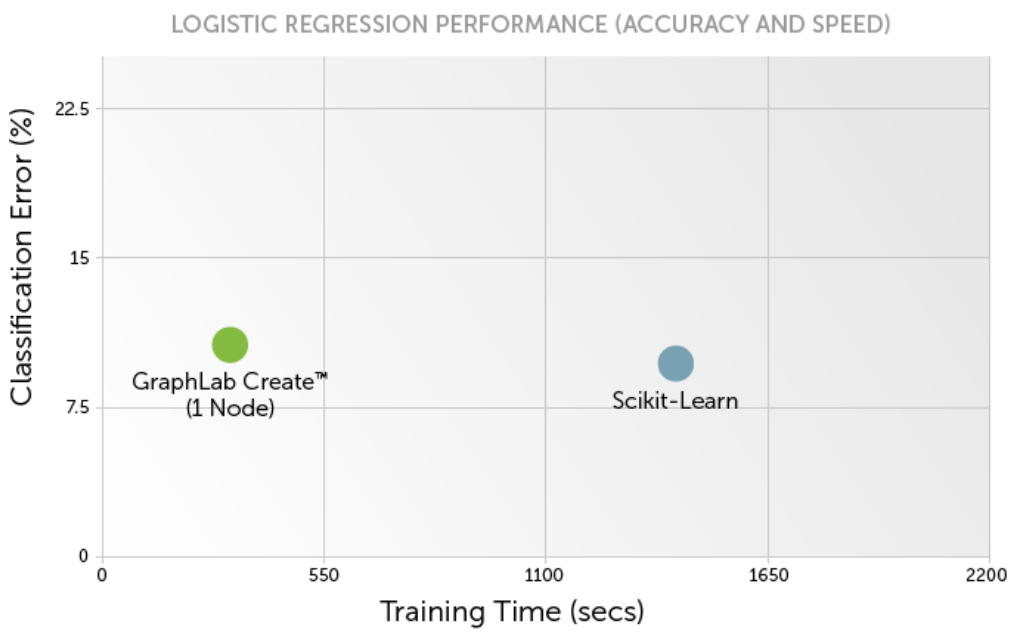
State-of-the-art solutions

- 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:

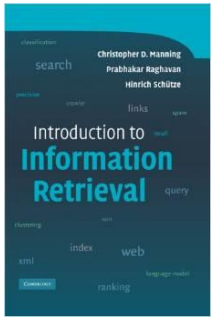
1. **Implicit feedback**
2. **Diverse and dynamic**

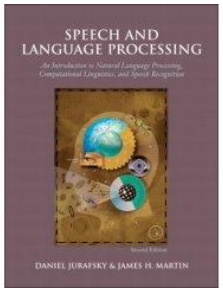
As data producer

Challenges:

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

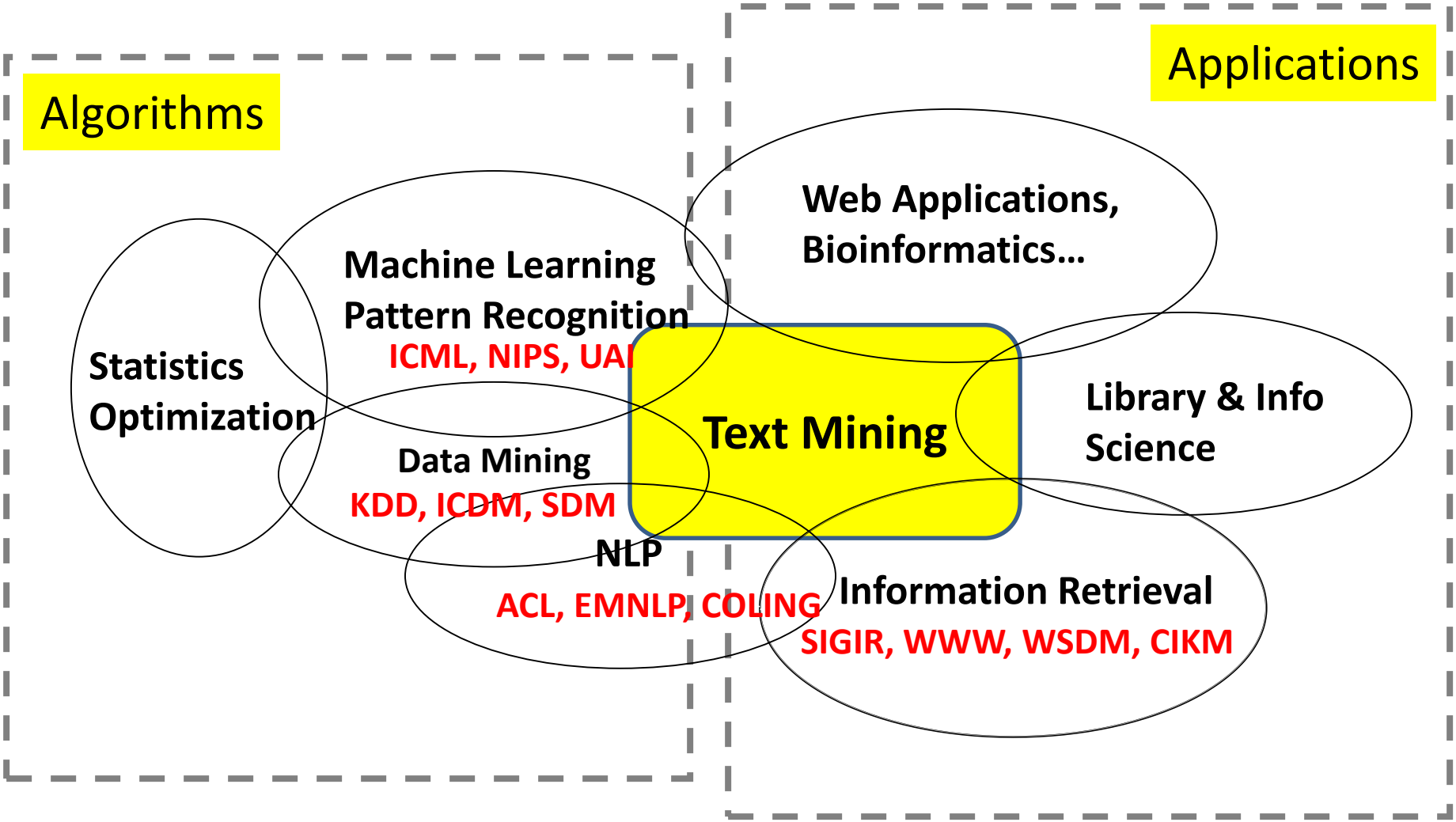
Text books





- ***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”!