**CIS 833 – Information Retrieval and Text Mining**

**Lecture 1**

# Introduction

August 25, 2015

Credits for slides: Hofmann, Mihalcea, Mobasher, Mooney, Schutze.

## Today

- Logistics
- What is this course about?
- What is Information Retrieval?

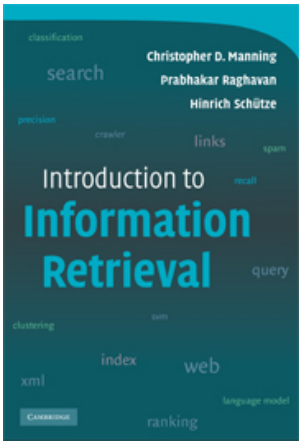
# Instructor

Dr. Doina Caragea
Office: 227C Nichols Hall
Phone: 785-532-7908
Email: dcaragea@ksu.edu

# Logistics

- **Meeting time**: TU 10-11:15am
- **Office hours**: Wed 1:30-2:30pm, or by appointment
- **Prerequisites**:
  - Basic knowledge on probability and statistics, data structures and algorithms.
  - Prior knowledge of Java.
  - The Yahoo! Hadoop implementation of MapReduce will be used for programming assignments. Prior experience with Hadoop is not required.
- Class materials available on K-State Online (Canvas)

# Recommended Textbook

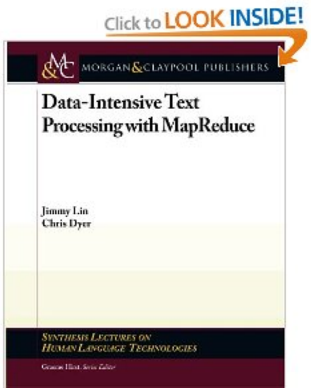


http://nlp.stanford.edu/IR-book/

# Topics

- Boolean retrieval
- The term vocabulary and postings lists
- Dictionaries and tolerant retrieval
- Index construction
- Index compression
- Scoring, term weighting and the vector space model
- Computing scores in a complete search system
- Evaluation in information retrieval
- Relevance feedback and query expansion
- XML retrieval
- Probabilistic information retrieval
- Language models for information retrieval
- Text classification and Naive Bayes
- Vector space classification
- Support vector machines and machine learning on documents
- Flat clustering
- Hierarchical clustering
- Matrix decompositions and latent semantic indexing
- Web search basics
- Web crawling and indexes
- Link analysis

# MapReduce Textbook



MORGAN & CLAYPOOL PUBLISHERS

Data-Intensive Text Processing with MapReduce

Jimmy Lin
Chris Dyer

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Graeme Hirst, Series Editor

- http://www.umiacs.umd.edu/~jimmylin/book.html

# Topics

- Introduction
- MapReduce Basics
- MapReduce algorithm design
- Inverted Indexing for Text Retrieval
- Graph Algorithms
- EM Algorithms for Text Processing
- Closing Remarks

# Course work and evaluation

- Individual assignments: 30% of grade
  - Theory exercises (15%)
  - Programming assignments (15%)

- Research review projects: 15% of the grade
  - Project proposal & presentation
  - Final project: report, presentation

- Two exams: each worth 25% of the grade

- Class participation: 5% of the grade

# Why Class Participation

"We learn:

10% of what we hear
30% of what we see
50% of what we see and hear
70% of what we discuss
80% of what we experience
95% of what we teach others."

> "Teach me and I will forget;
> show me and I may remember;
> involve me and I will understand."

~ Chinese Proverb.

## Lecture structure

- Short review
- New concepts
- Exercises
- Conclusions
  - Evaluate the lecture, point out sections that are unclear, make suggestions for improvements
  - Do not save content questions for the end of the class. Ask them in class.

## Submission policies

- Programming assignments will be submitted through K-State online.
- Homework assignments are due one week after they are assigned (unless otherwise noted).
- Late submissions are not encouraged. I will accept late submissions at my discretion, but there might be grading penalties.

## Collaboration policies

- Students are encouraged to discuss the course material, concepts, and assignments, but they should write their answers independently.
- For each assignment, you are required to list students with whom you have discussed the assignment.
- Your submission should reflect your own knowledge and you should be able to reproduce the material you turn in at any time.
- Sharing answers will not be tolerated.
- Plagiarism will not be tolerated either.
- Appropriate citations for any external sources used in your work are mandatory. Never use sentences or phrases taken directly from a paper you are reviewing.

## Honor Pledge

Honor Pledge applies to all assignments, examinations, and other course work undertaken by students:

“On my honor, as a student, I have neither given nor received unauthorized aid on this academic work.”

## Students with disabilities

"Any student with a disability who needs an accommodation or other assistance in this course should make an appointment to speak with me as soon as possible."

## Other policies

- Expectations for classroom conduct:
  - “All student activities in the University, including this course, are governed by the Student Judicial Conduct Code as outlined in the Student Governing Association By Laws, Article VI, Section 3, number 2. Students who engage in behavior that disrupts the learning environment may be asked to leave the class.”
- No make-up exams
- No incompletes

## Who are you?

- Background?
- Research interests?
- Expectations from CIS 833?
- Something fun/interesting about you ☺

## What is this course about?

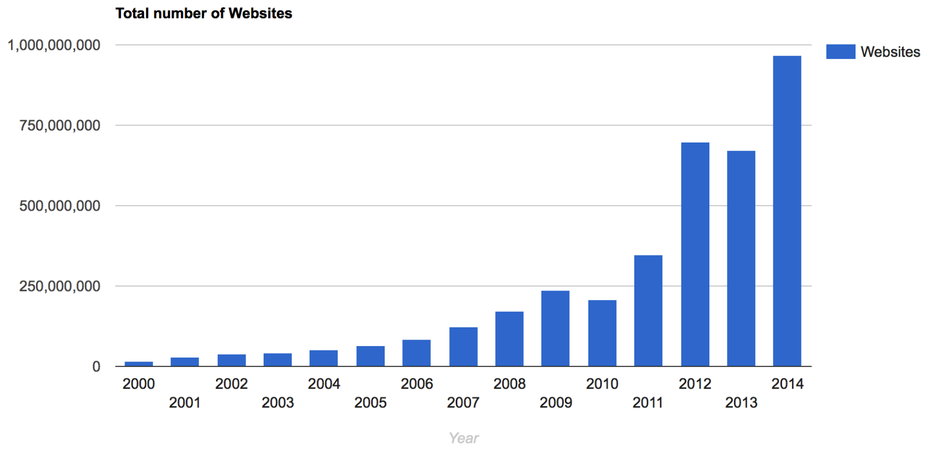
- Processing
- Indexing
- Retrieving

… textual data

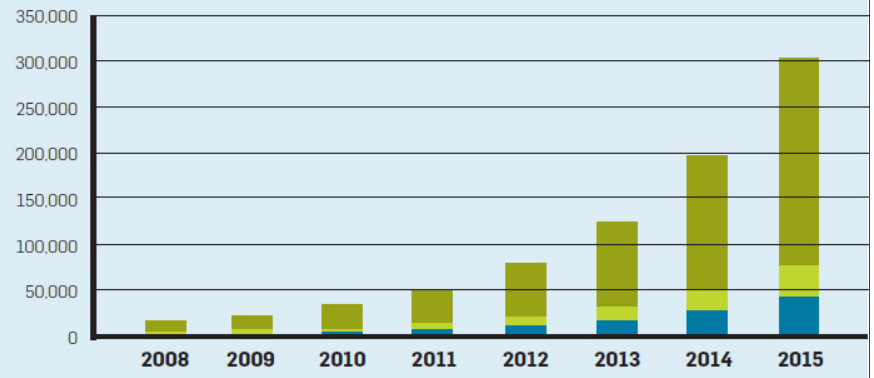
## Need for IR?

# Large digital information repositories

- World Wide Web (~ 1,000,000,000 websites)
- Digital Libraries (e.g., California Digital Library)
- Special purpose content providers (e.g., Lexis Nexis)
- Company intranets and digital assets
- Scientific literature libraries (e.g., CiteSeer)
- Medical information portals (e.g., MedlinePlus)
- Patent databases (e.g., US Patent Office)
- Online encyclopedias (e.g., Wikipedia)



**Internet Live Stats** (http://www.internetlivestats.com/total-number-of-websites/#sources)

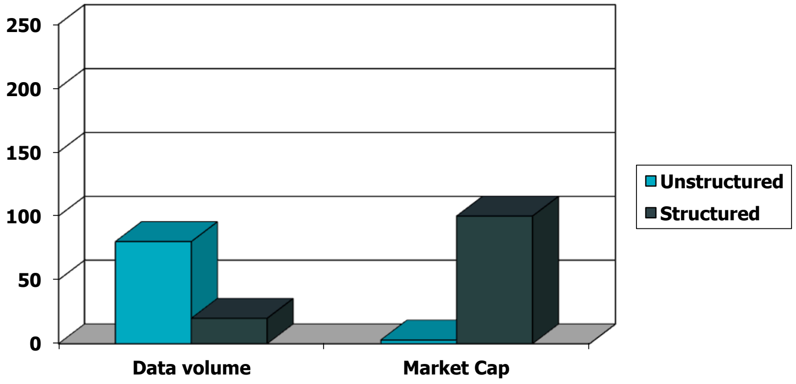


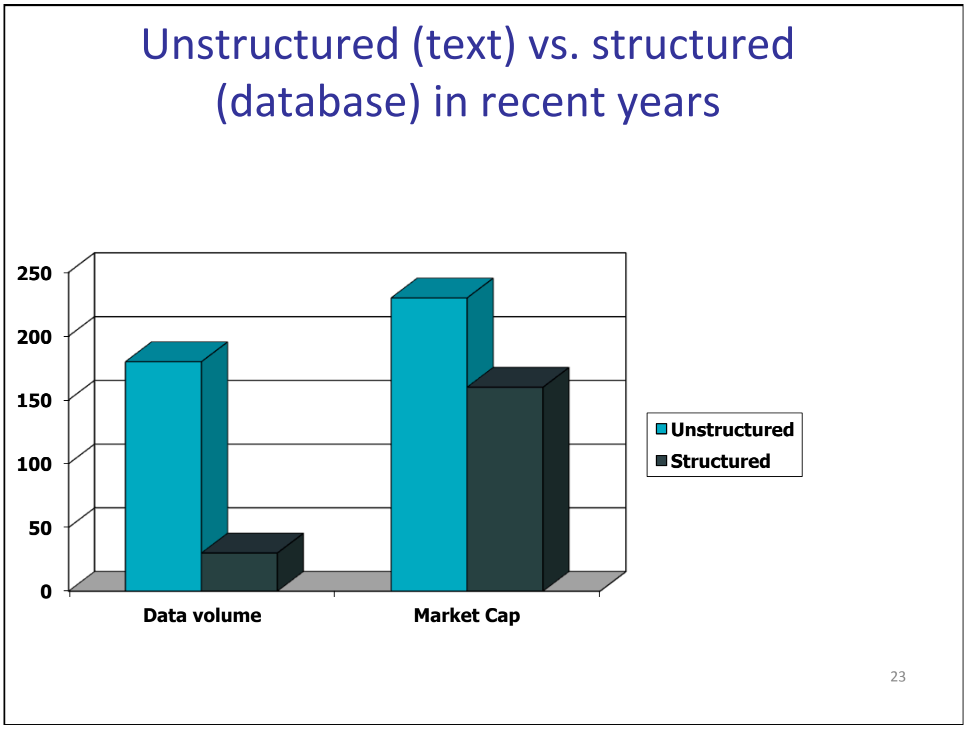
Total archived capacity, by content type, worldwide, 2008–2015 (petabytes)

| | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|
| 🟩 Unstructured | 11,430 | 16,737 | 25,127 | 39,237 | 59,600 | 92,536 | 147,885 | 226,716 |
| 🟩 Database | 1,952 | 2,782 | 4,065 | 6,179 | 9,140 | 13,824 | 21,532 | 32,188 |
| 🟦 Email | 1,652 | 2,552 | 4,025 | 6,575 | 10,411 | 16,796 | 27,817 | 44,091 |

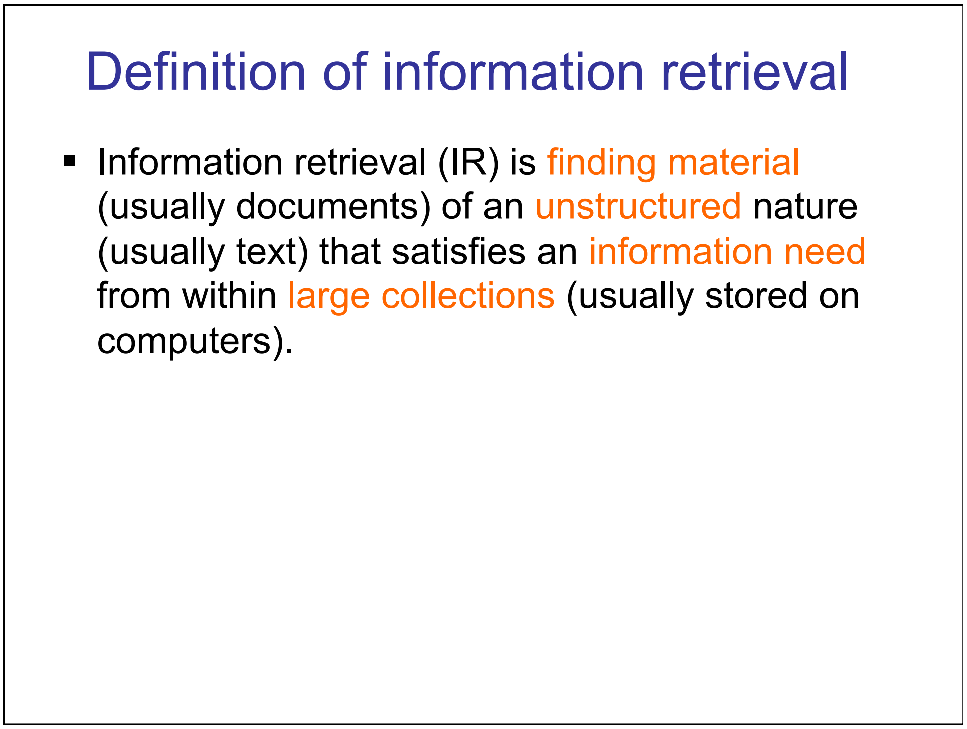
Communications of the ACM (http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext)

## Unstructured (text) vs. structured (database) data in the mid-nineties





## Unstructured (text) vs. structured (database) in recent years



## Definition of information retrieval

- Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

## Various needs for information

- Search for documents that fall in a given topic
- Search for specific information
- Search for an answer to a question
- Search for information in a different language
- …
- Search for images
- Search for music
- Search for a (candidate) friend

## Other definitions of IR

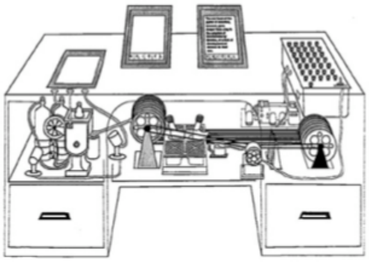
**Salton (1989):** Information-retrieval systems process files of records and requests for information, and identify and retrieve from the files certain records in response to the information requests. The retrieval of particular records depends on the similarity between the records and the queries, which in turn is measured by comparing the values of certain attributes to records and information requests.

**Fuhr (1991):** IR deals with *uncertainty* and *vagueness* in information systems

- *uncertainty*: available representation does typically not reflect true semantics/meaning of objects (text, images, video, etc.)
- *vagueness*: information need of user lacks clarity, is only vaguely expressed in query, feedback or user actions

# Memex

- It is not enough to store information!
- The idea of an easily accessible, individually configurable storehouse of knowledge - beginning of the literature on mechanized information retrieval:



"Consider a future device for individual use, which is sort of mechanized private file and library. It needs a name, and to coin one at random, '**memex**' **will do.** A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Vannevar Bush, *As we may think*, Atlantic Monthly, 176 (1945), pp.101-108

http://www.theatlantic.com/doc/194507/bush

# Search as a Principle & Problem

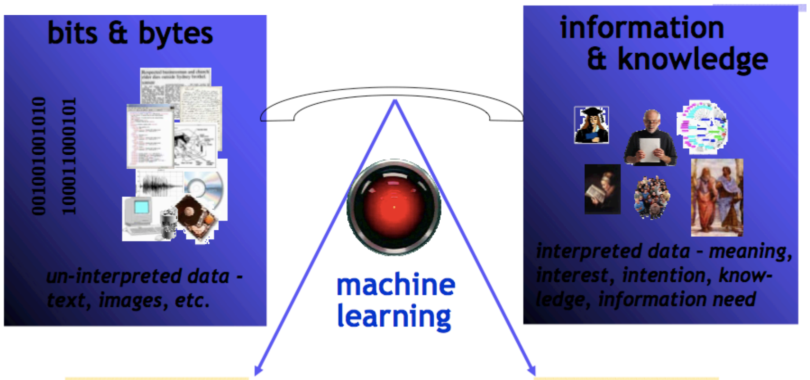
"The difficulty seems to be, not so much that we public unduly in view of the extent and variety of present day interests, but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships."

V. Bush*, As we may think,* Atlantic Monthly, 176 (1945), pp.101-108

We live in a search society – belief that (almost) everything is known, we just have to find the information

We search for everything –the right book, movie, car, house, vacation trip, bargain, partner, search engine etc.

## Machine Learning in IR


bits & bytes
0010010010101
100011000101
un-interpreted data - text, images, etc.
information & knowledge
interpreted data - meaning, interest, intention, knowledge, information need
machine learning

**interpretation**

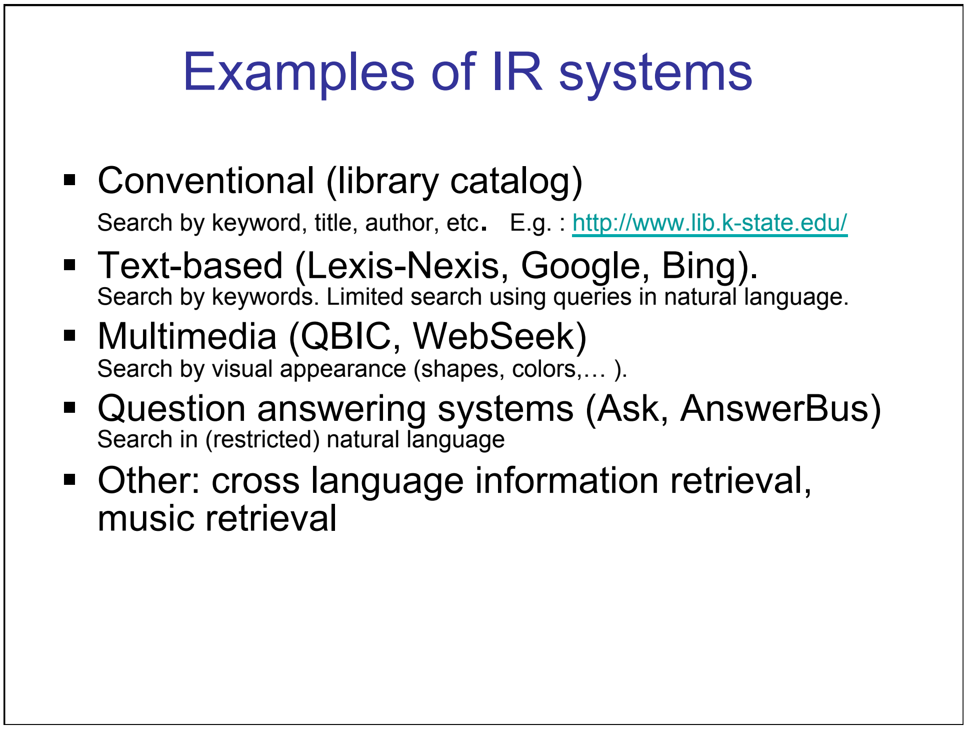
***unsupervised learning & data mining:*** *discover hidden regularities, generate semantically meaningful representations, predictive modeling, statistics*

**generalization**

***supervised learning:*** *generalize from given examples, classification & recognition, emulate human experts*

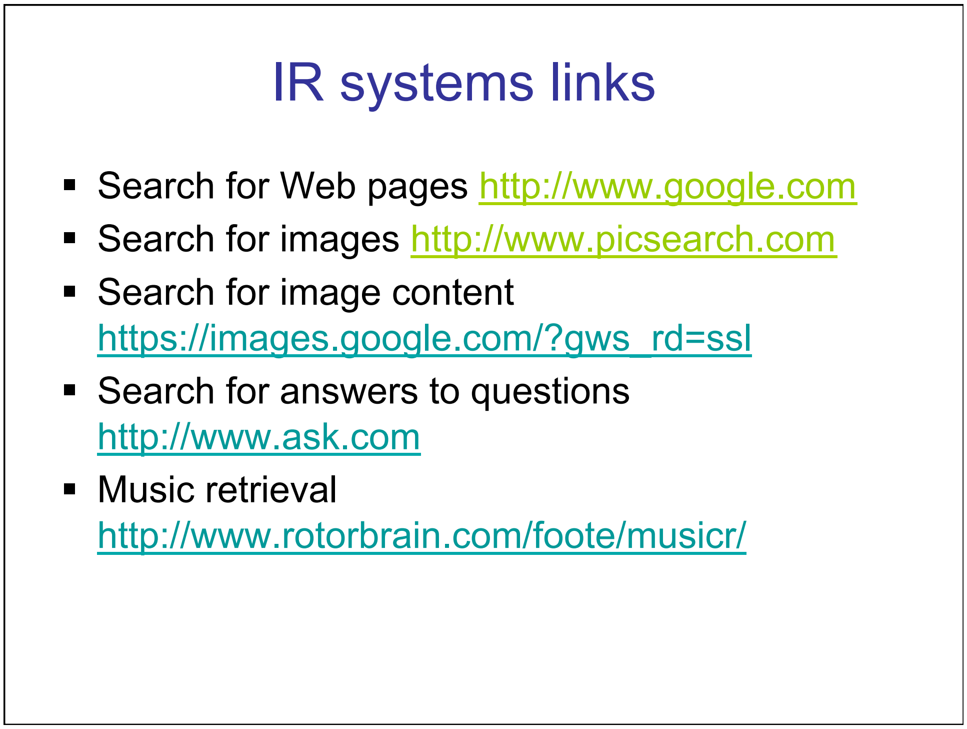
## Machine Learning: some IR directions

- Text Clustering
  - Clustering of IR query results
  - Automatic formation of hierarchies
- Text Categorization
  - Automatic hierarchical classification (Yahoo)
  - Adaptive filtering/routing/recommending
  - Automated spam filtering.
- Learning for Information Extraction
- Text Mining



## Examples of IR systems

- Conventional (library catalog)
  Search by keyword, title, author, etc. E.g. : http://www.lib.k-state.edu/
- Text-based (Lexis-Nexis, Google, Bing).
  Search by keywords. Limited search using queries in natural language.
- Multimedia (QBIC, WebSeek)
  Search by visual appearance (shapes, colors,… ).
- Question answering systems (Ask, AnswerBus)
  Search in (restricted) natural language
- Other: cross language information retrieval, music retrieval



## IR systems links

- Search for Web pages http://www.google.com
- Search for images http://www.picsearch.com
- Search for image content https://images.google.com/?gws_rd=ssl
- Search for answers to questions http://www.ask.com
- Music retrieval http://www.rotorbrain.com/foote/musicr/

## Information Retrieval

- The processing, indexing and retrieval of textual documents.
- Searching for pages on the World Wide Web is the most recent “killer app.”
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.

## In this course, we ask…

- What makes a system like Google, Yahoo! or Bing tick?
  - How does it gather information?
  - What tricks does it use?
  - Extending beyond the Web
- How can those approaches be made better?
  - Natural language understanding?
  - User interactions?
- What can we do to make things work quickly?
  - Faster computers?
  - Caching?
  - Compression?
- How do we decide whether it works well?
  - For all queries?  For special types of queries?
  - On every collection of information?
- What else can we do with the same approach?
  - Other media?
  - Other tasks?

## Before that…

We will learn about MapReduce framework.

## Reading for next time

- Chapters 1 from MapReduce textbook