4840-1031: Monday 14:55-16:40
Web Engineering (ウェブ工学）
https://tinyurl.com/webeng21

Masashi Toyoda

E-mail: mtoyoda@acm.org

The Web as a Platform

- Sharing information
  - Docs, news, images, movies, files…
- Mass media
  - TV, movie, newspaper, radio, magazine…
- Services
  - E-commerce, trading, gaming…

The Web as Data Sources

- Texts
- Images, movies
- Graphs
  - Hyperlink, social network, P2P network…
- Geospatial data
  - Map queries, check-in…
- Time series
  - User traffic, sensor monitoring…

Goal

- Learn recent trends in Web related research including
  - Search engine, Information Retrieval
  - Natural Language Processing
  - Machine Learning, Data Mining
  - Social Media/Network
  - Big Data
  - Cyber-Physical Systems
  - Internet of Things (IoT)

Format

- Seminar Style (Online)
- Three or four students present research papers (published on top conferences/journals) at each session
  - Presentation: 15 min.
  - Q&A: 10 min.

Organization of sessions

- First half (Session 1 – 4)
  - Test of time award papers from WWW, KDD, etc.
- Second half (Session 5 – 11)
  - Recent research papers
    - Web: WWW, WSDM
    - Database: SIGMOD, VLDB, ICDE
    - IR: SIGIR
    - Data mining: KDD, ICDM
    - AI: AAAI, ICWSM
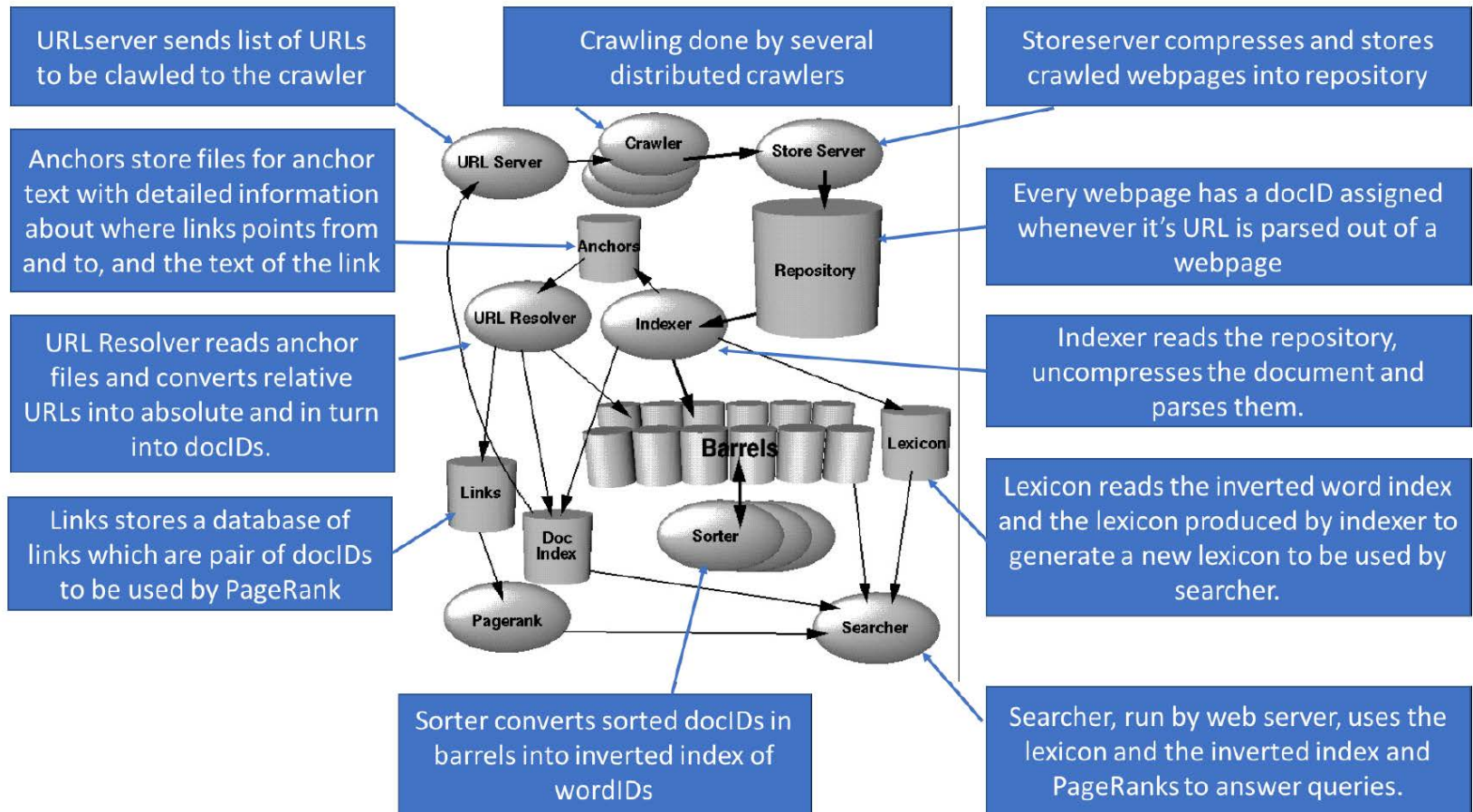   - NLP: ACL, EMNLP
    - Etc.

Test of Time Award Papers

- If you want to present in the first half sessions, you need to choose a test of time award paper in WWW, KDD, SIGIR, etc.
  - WWW Seoul Test of Time Awards https://www.iw3c2.org/ToT/
  - KDD Test of Time Awards http://www.kdd.org/awards/kdd-test-of-time-award
  - SIGIR Test of Time Awards http://sigir.org/awards/test-of-time-awards/

WWW2015: Test of Time Awards
The Anatomy of a Large-Scale Hypertextual Web
Search Engine [S. Brin, L. Page, WWW1998]

• Google's architecture in the early days

URLserver sends list of URLs
to be clawled to the crawler

Crawling done by several
distributed crawlers

Storeserver compresses and stores
crawled webpages into repository

Anchors store files for anchor
text with detailed information
about where links points from
and to, and the text of the link

Every webpage has a docID assigned
whenever it's URL is parsed out of a
webpage

URL Resolver reads anchor
files and converts relative
URLs into absolute and in turn
into docIDs.

Indexer reads the repository,
uncompresses the document and
parses them.

Links stores a database of
links which are pair of docIDs
to be used by PageRank

Lexicon reads the inverted word index
and the lexicon produced by indexer to
generate a new lexicon to be used by
searcher.

Sorter converts sorted docIDs in
barrels into inverted index of
wordIDs

Searcher, run by web server, uses the
lexicon and the inverted index and
PageRanks to answer queries.

URL Server, Crawler, Store Server, Anchors, Repository, URL Resolver, Indexer, Barrels, Lexicon, Links, Doc Index, Sorter, Pagerank, Searcher

WWW2016: Test of Time Awards
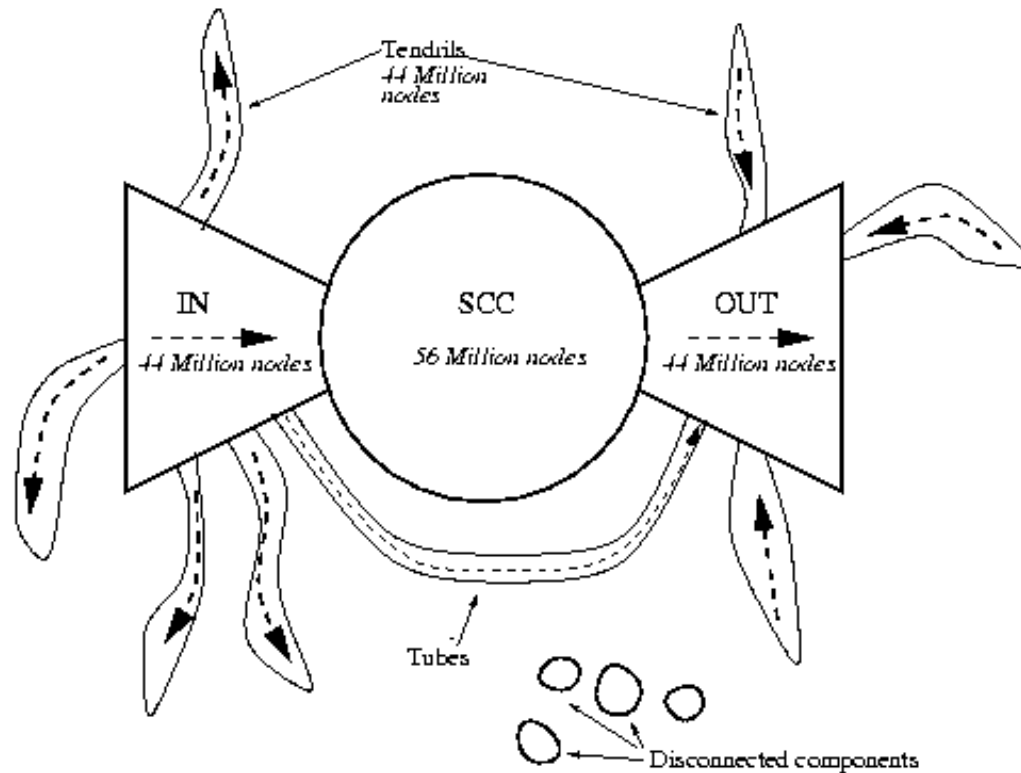Item-based collaborative filtering recommendation algorithms [B. Sarwar+, WWW1998]

- Many collaborative filtering systems now use this approach
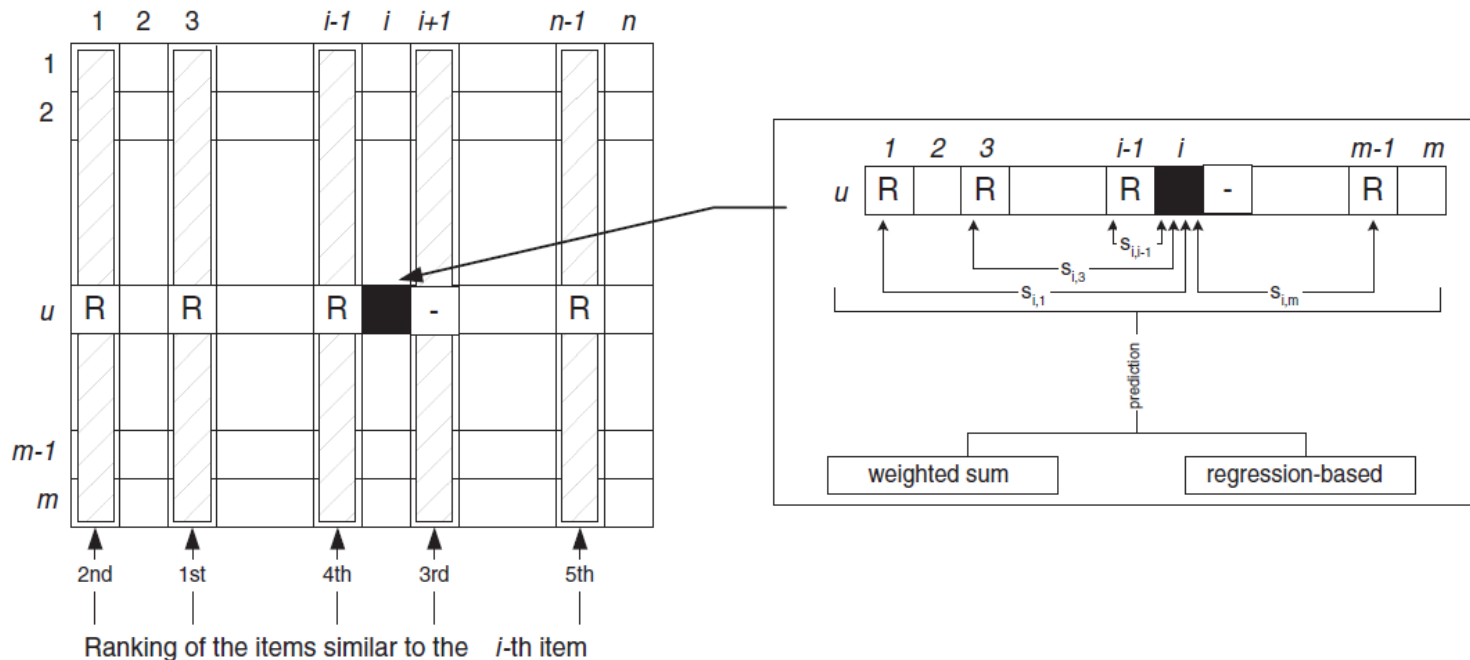
WWW2017: Test of Time Awards
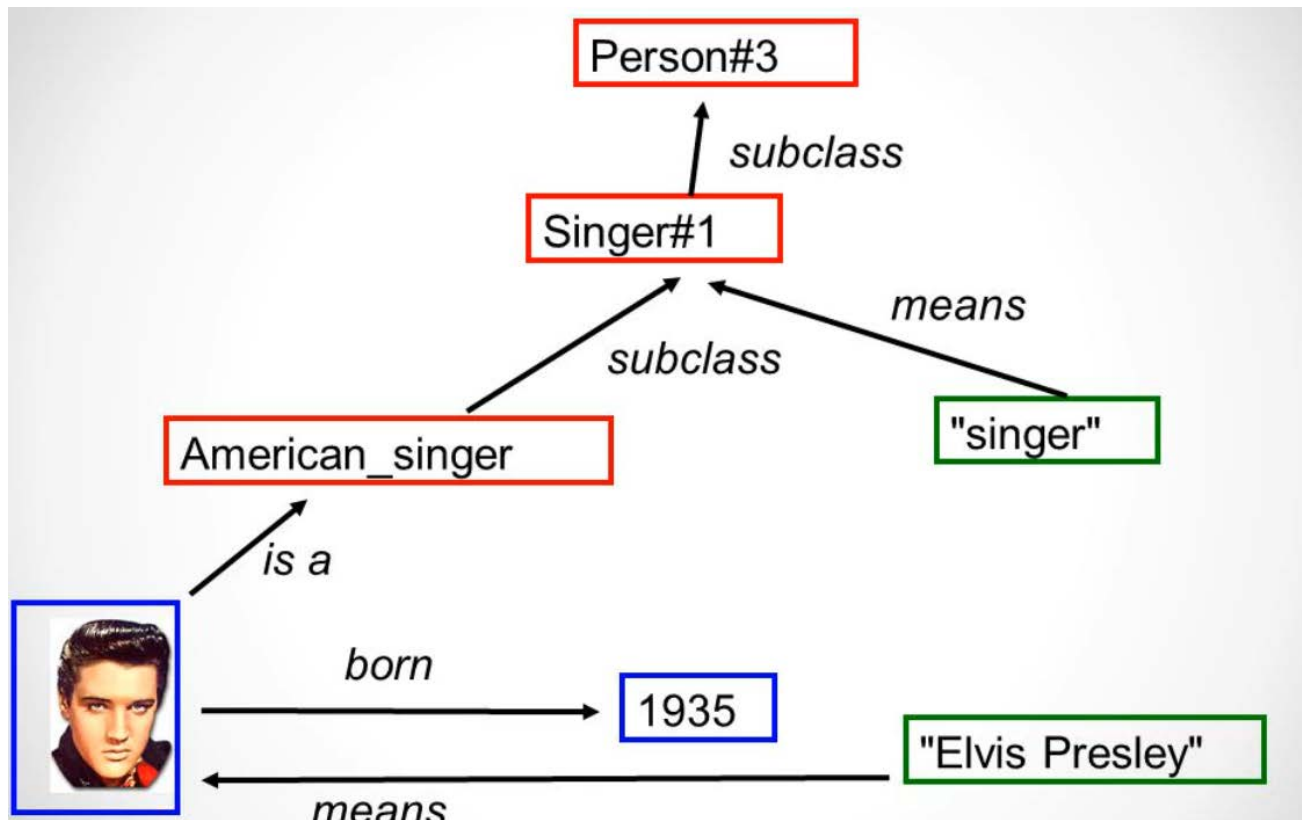Graph Structure in the Web [A. Broder+, WWW2000]

- The first large-scale empirical study of the dynamics of the Web

WWW2018: Test of Time Awards
YAGO: A Core of Semantic Knowledge Unifying
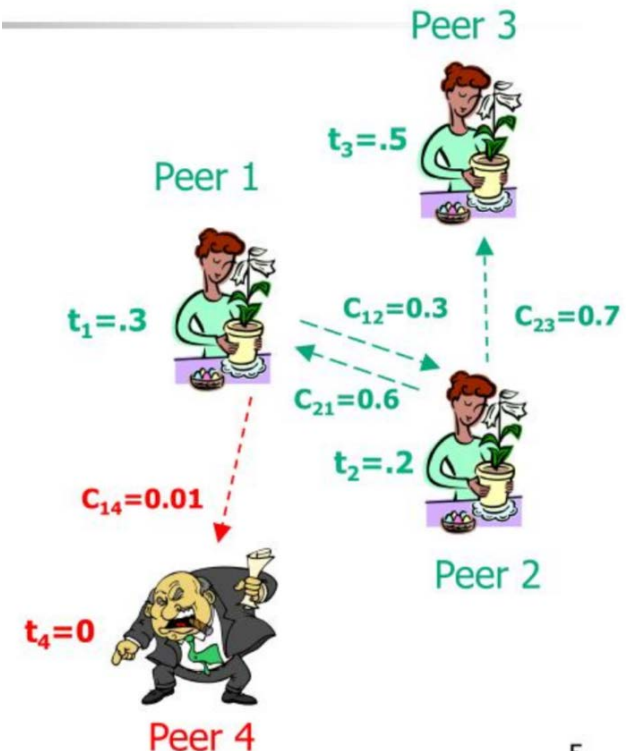WordNet and Wikipedia [F. Suchanek+, WWW2007]

- One of the first projects to extract semantic
  knowledge graph at large scale from Wikipedia

WWW2019: Test of Time Awards
The EigenTrust Algorithm for Reputation Management in P2P Networks [Kamvar+, WWW2003]

- In P2P file sharing networks, reduce inauthentic files distributed by malicious peers
  - Goal: Identify sources of inauthentic files by gibing each peer a trust value based on its previous behavior
  - Algorithm: Propagate local trust values to friends, friends of friends, and so on
    - Propose a distributed algorithm to calculate this eigen trust value

Peer 3

$t_3 = .5$

Peer 1

$t_1 = .3$

$C_{12} = 0.3$ $C_{23} = 0.7$

$C_{21} = 0.6$

$t_2 = .2$

$C_{14} = 0.01$

$t_4 = 0$

Peer 2

Peer 4

WWW2020: Test of Time Awards
Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews [Dave+, WWW2003]

- Aggregate product reviews in various sites
  - Extract product attributes and aggregate pos/neg about them

WWW2021: Test of Time Awards
Information Credibility on Twitter [Castillo+, WWW2011]

- Classify tweets related to newsworthy events by their **believability**
  - Users believe those tweets are to be true or false
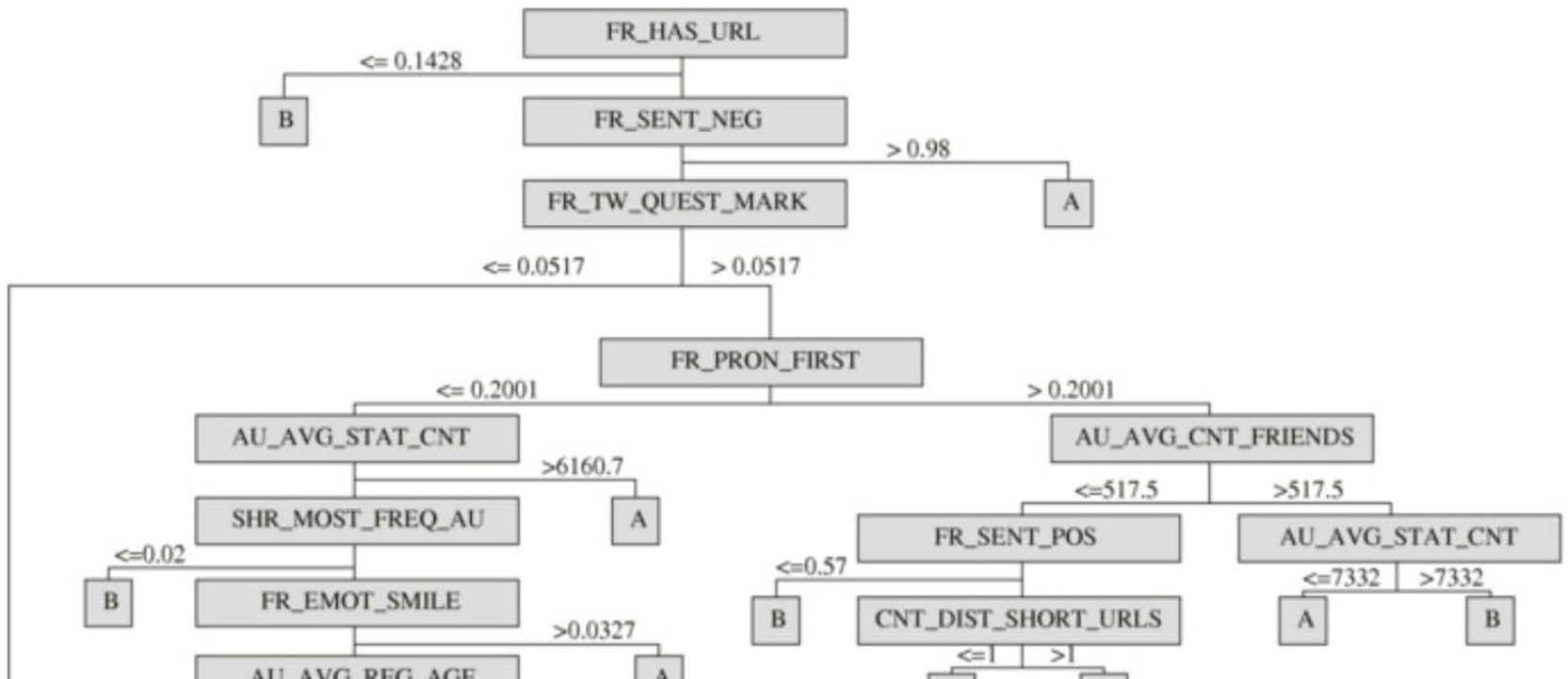
Recent research papers

- If you want to present in the last half sessions, you need to choose:
  - Choose full research papers (over 8 pages) in top conferences/journals published in recent 4 years (2018 – 2021)
  - **DO NOT** choose workshop, poster, and preprint (e.g. arXiv.org) papers
  - Topics of the papers should be related to Web, and include experiments with real data
  - Cross-disciplinary papers are recommended

Presentation

- Talk: 15 min., QA: 10 min.
- Language: Japanese or English
- Prepare slides including the following:
  - Why did you choose the paper?
  - Importance of the paper
  - Motivation, purpose, and goal
  - Novelty (compared with related work)
  - Method
  - Experimental results
  - Strong points and weak points

Slack

- We use slack.com to announcement and chat in sessions
- I will check your questions and comments in sessions and mark them if they are "good" (Mark will become scores for your credits)

- After receiving your e-mail, I will invite you to **webeng2021.slack.com**
- Your name in slack should be "Your_Name_StudentID" (Please do no use nicknames)

Credits

- Point system:
  - **100** points (優)
  - **90** points (良)
  - **70** points (可)
  - Less than 70 points (不可)
- Scores:
  - Presentation: **40** pts.
  - Good question/comment: **10** pts./session (Do not count multiple times in a session)
  - Attendance: **0** pts.

Schedule

Date		Session
10/4	Mon	Guidance
10/11	Mon	Guidance
10/18	Mon	Session 1: Test of time award papers
10/25	Mon	Session 2: Test of time award papers
11/1	Mon	Session 3: Test of time award papers
11/8	Mon	Session 4: Test of time award papers
11/15	Mon	Session 5: Recent papers
11/22	Mon	Canceled (Komaba Festibal)
11/29	Mon	Session 6: Recent papers
12/6	Mon	Session 7: Recent papers
12/13	Mon	Session 8: Recent papers
12/20	Mon	Session 9: Recent papers
12/27	Mon	Session 10: Recent papers
1/17	Mon	Session 11: Recent papers
1/24	Mon	Session 12: Recent papers

Assignment

- Today, we decide presenters for the 1st and 2nd sessions
- All students should e-mail following info. **by 15 Oct**.
  - Student ID Number, Name, e-mail for slack
  - 3 candidate dates for presentation, and
  - 3 papers you want to present with complete bib info.
    - Authors, Title, Conference/Journal, pages, year
- Assignment & Schedule will be released **on 18 Oct**.
- Note:
  - Students who select earlier day have more chance to present a preferred paper
  - Let me know ASAP, if you change the paper to present