# EE412 Introduction to Big Data

Project 2 tutorial

# Introduction

- Topics: Collaborative filtering
  - Matrix completion
- Goal
  - Predict user's hidden ratings
- Training dataset
  - User ID
  - Item ID
  - Ratings on items
  - Timestamps when user rate the item

# Data format

- User IDs and item IDs are integer number
- Ratings $\in \{1,2,3,4,5\}$
- Timestamp in Unix time

| User Id | Item id | rating | timestamp |
|---|---|---|---|
| 1 | 1 | 2 | 1260759144 |
| 1 | 32 | 4 | 1260759205 |
| … | | … | … |
| 7 | 145 | 5 | 851868044 |

# Tasks

- Given user's rating on some item, you should predict all the hidden ratings of each users

- Performance measure: RMSE

# Guide

- You can use any algorithms
- But recommend to begin with what you've learned in the class

- Split your training data to train/val set
  - Validate your algorithm with validation set
  - Check how effective each modifications you've made

# Implementation

- You can use Matlab or Python for this project
- Submitted code should include
  - Training code (code to reproduce your model)
  - Evaluation code (making output result file)
- Your result should be reproducible on TA's machine with your code

# Submission

- Due date: 18th Dec, 11:59 PM
- Submit to KLMS
- You should submit
  - Source code
  - Documentation
  - Result file
- Submit source code, documentation, and result file with zip file
  - file name: **[student-id].zip**
- Documentation
  - Explanation about your method
  - Explanation about your implementation
  - Explanation about how to run your code
  - Reference if you have any
  - file name: **doc_[student-id].xx** (pdf, docs, **No** hwp)

# Submission

- Result file
  - Result on **test set**
  - Format
    - Each line contains one element of matrix
    - <user_id><*comma*><item_number><*comma*><ratings>
      - Example:

```
1,1,2.5
1,2,5
…
2,40,4.5
…
7000,234,3.5
7000,235,0.5
```

# Evaluation

- Accuracy (40%)
  - **0 if submitted algorithm fails to run.**
  - **0 if result has invalid value.**
- Novelty (50%)
  - 0 if you use open-source implementation without any modification
- Documentation (10%)

# Announcement

- Make a team
  - 1~3 members / team
  - If you have mates, please email to TA (until this Sunday, 26th Nov)
  - If you want to do it alone or same team as in Project 1, you don't have to send e-mail
    - If you don't send me an e-mail, I'll consider that you work as same team (or do it alone if you did project #1 alone) in project #1
- This tutorial and dataset will be posted today.
- If you have any questions, please contact to yh.jang@kaist.ac.kr