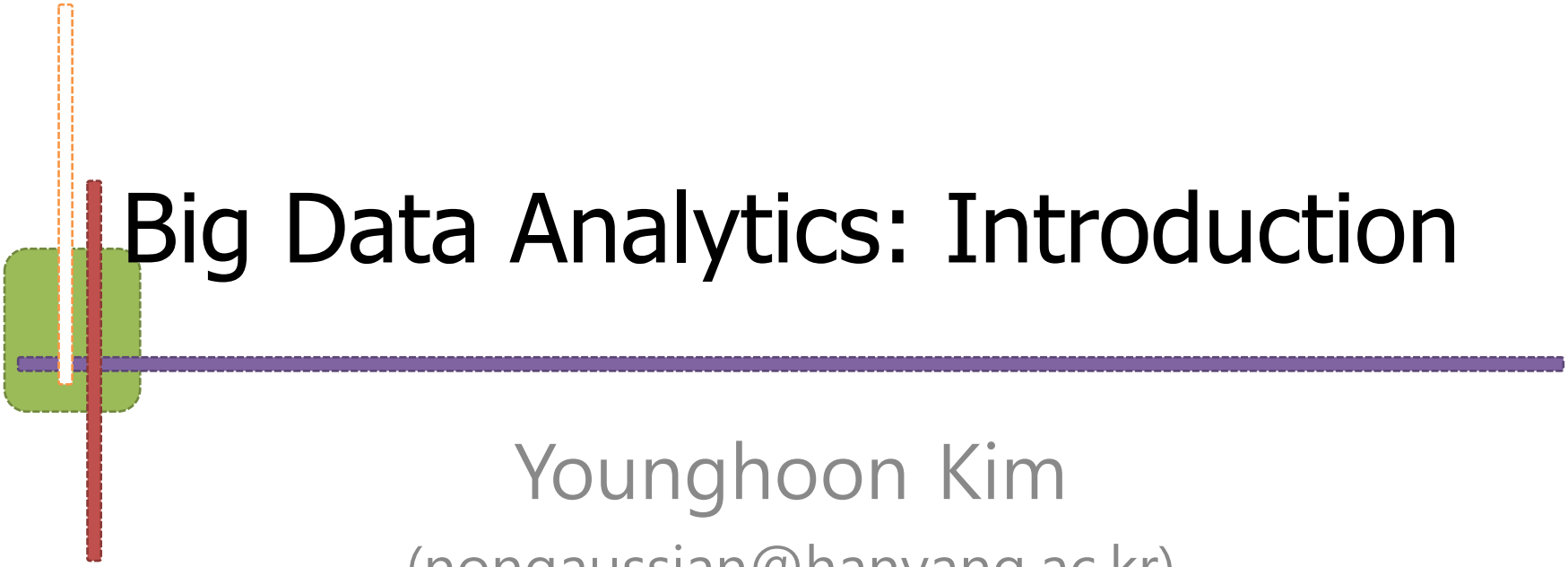


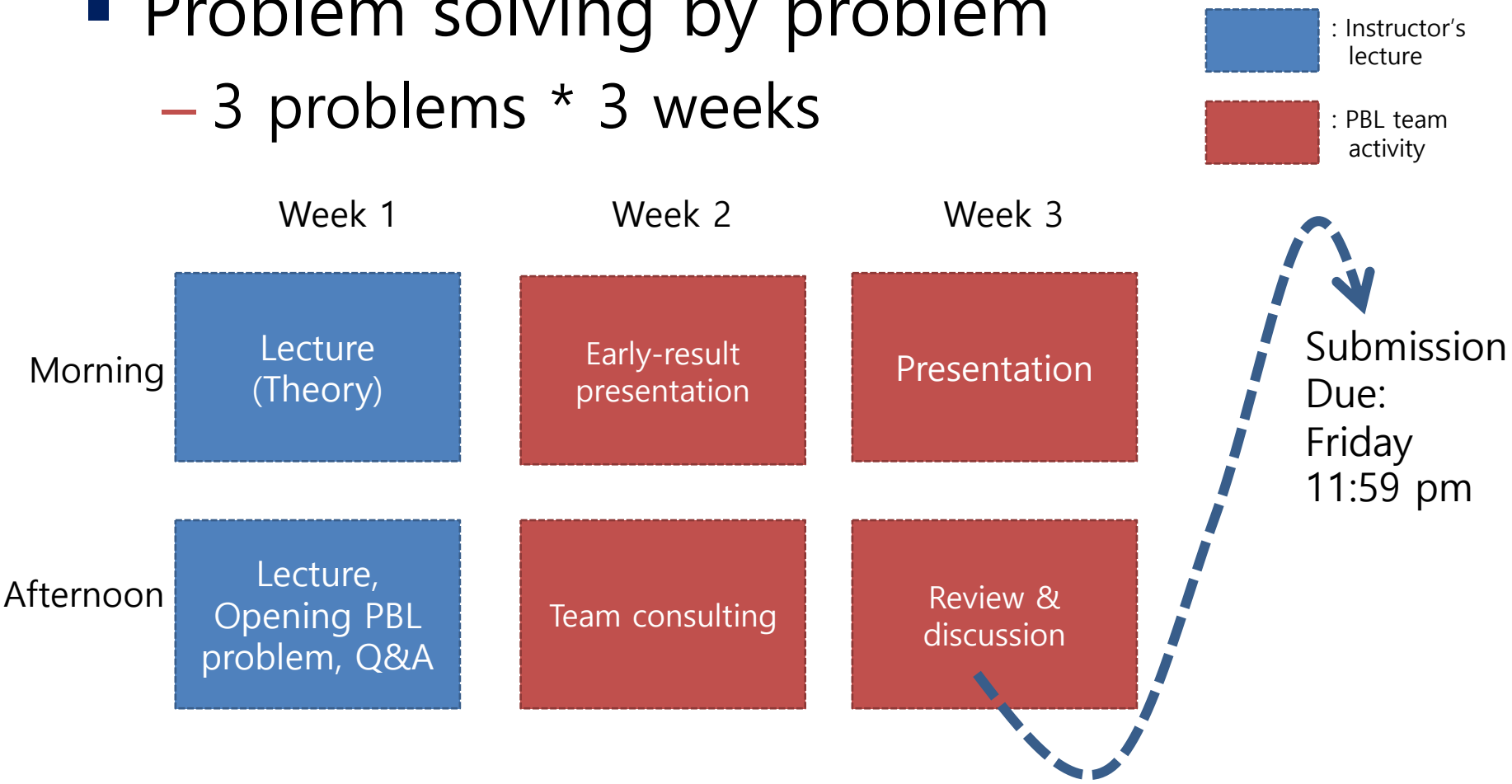
Big Data Analytics: Introduction

Younghoon Kim

(nongaussian@hanyang.ac.kr)

PBL Class

- Problem solving by problem
  - 3 problems * 3 weeks



Outputs

- Submissions
  - For every problem, team's leader must submit
    - Source code & prediction result
    - Early-result presentation slides
      - Including problem analysis, team's strategy, temporary result and any trouble
    - Final presentation slides
      - Including the details of techniques used, performance result and discussion
  - Final
    - Term report (1 page) (individual; 개인별 1건)
      - What you have learn and any suggestion for improving class
    - Contribution report (per team; 팀별 1건)
      - Contribution of each member
      - Purpose: to find free-riders

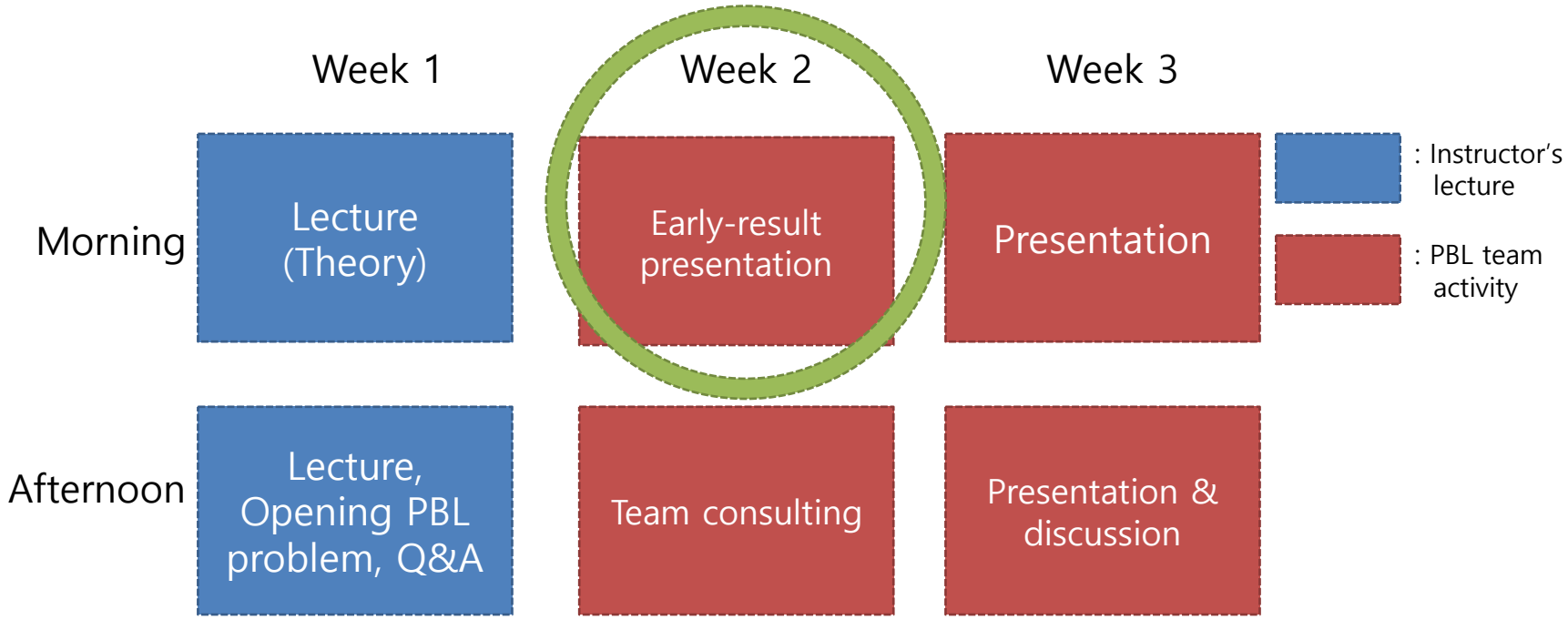
Tools

- Evaluation machine
  - Ubuntu 16.04
  - Java (recent version)
- Language
  - Python, Java, Scala
  - Statistics tools such as SPSS, SAS and matlab are not allowed
- Library
  - Spark ~~2.2.0~~
  - Scikit-learn ~~0.18.2~~
  - TensorFlow ~~1.4.0~~
  - Any external libs are allowed, but should be included in your submission, and TA can run with one line of commands (you cannot request installation of extra packages)

Team

- Find your team
  - Make a team with members you want to work together (until the next week class)
  - Then, choose your team's leader
  - Team size = 3 ~ 5 (if you have any trouble to make a team, TA will help you)
- Collaboration of not only inter-team but also intra-team is very important!

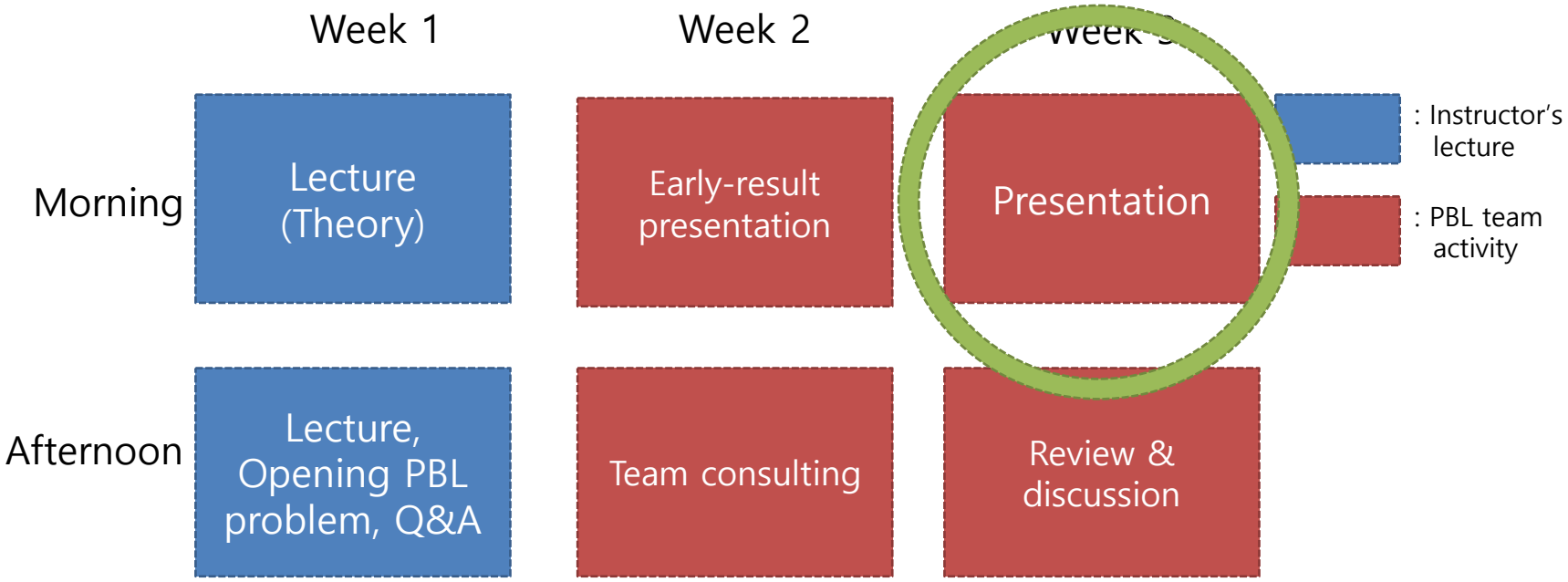
Early-result Presentation (10min)



Early-result Presentation (10min)

- Share your team's ideas with other teams!
- Provide
  - Your observation from data
  - How to preprecess the raw data
  - Strategy to tackle the problem
  - (if possible) temporary result

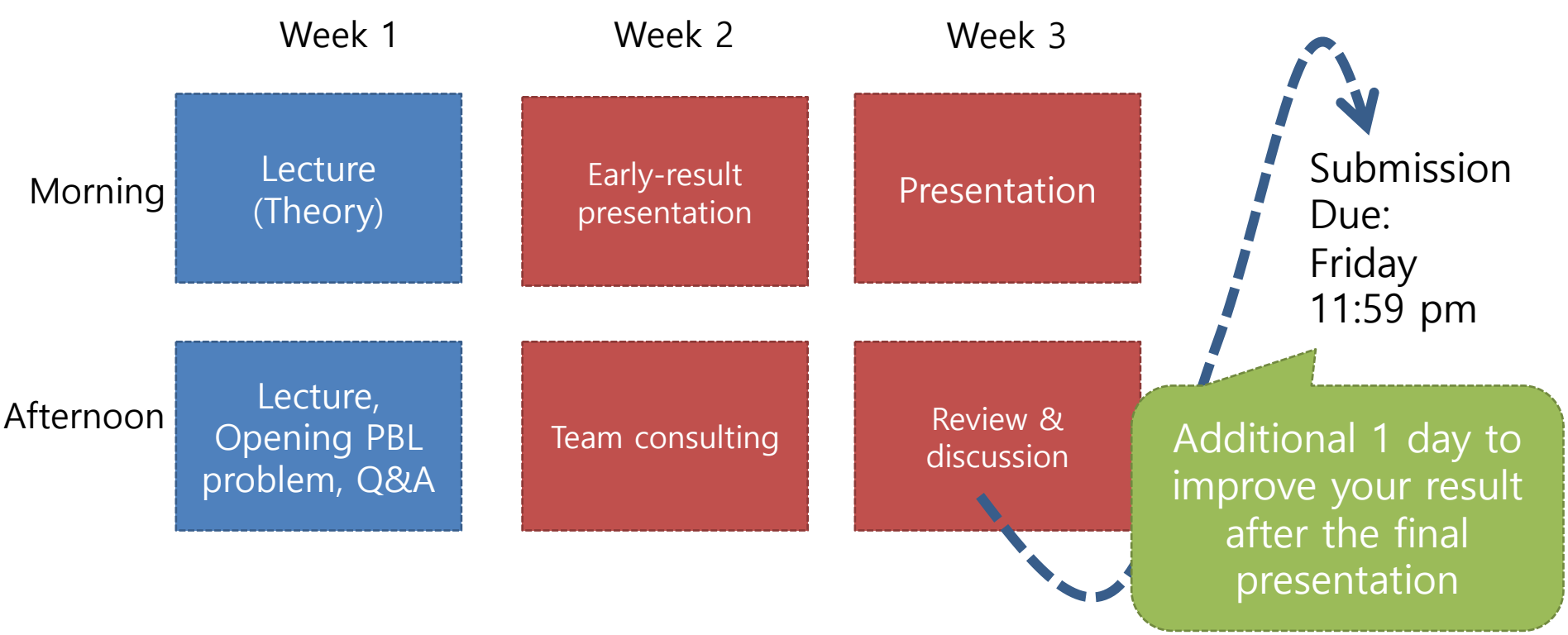
Final Presentation (15min)



Final Presentation

- Show
  - Test result
  - Your key idea & algorithms (don't repeat something already discussed by instructor)
  - Discussion – boast of how you obtain such WONDERFUL result
- Peer-to-peer grading
  - Grade the other teams
  - Presentation score / technique score

Final Submission

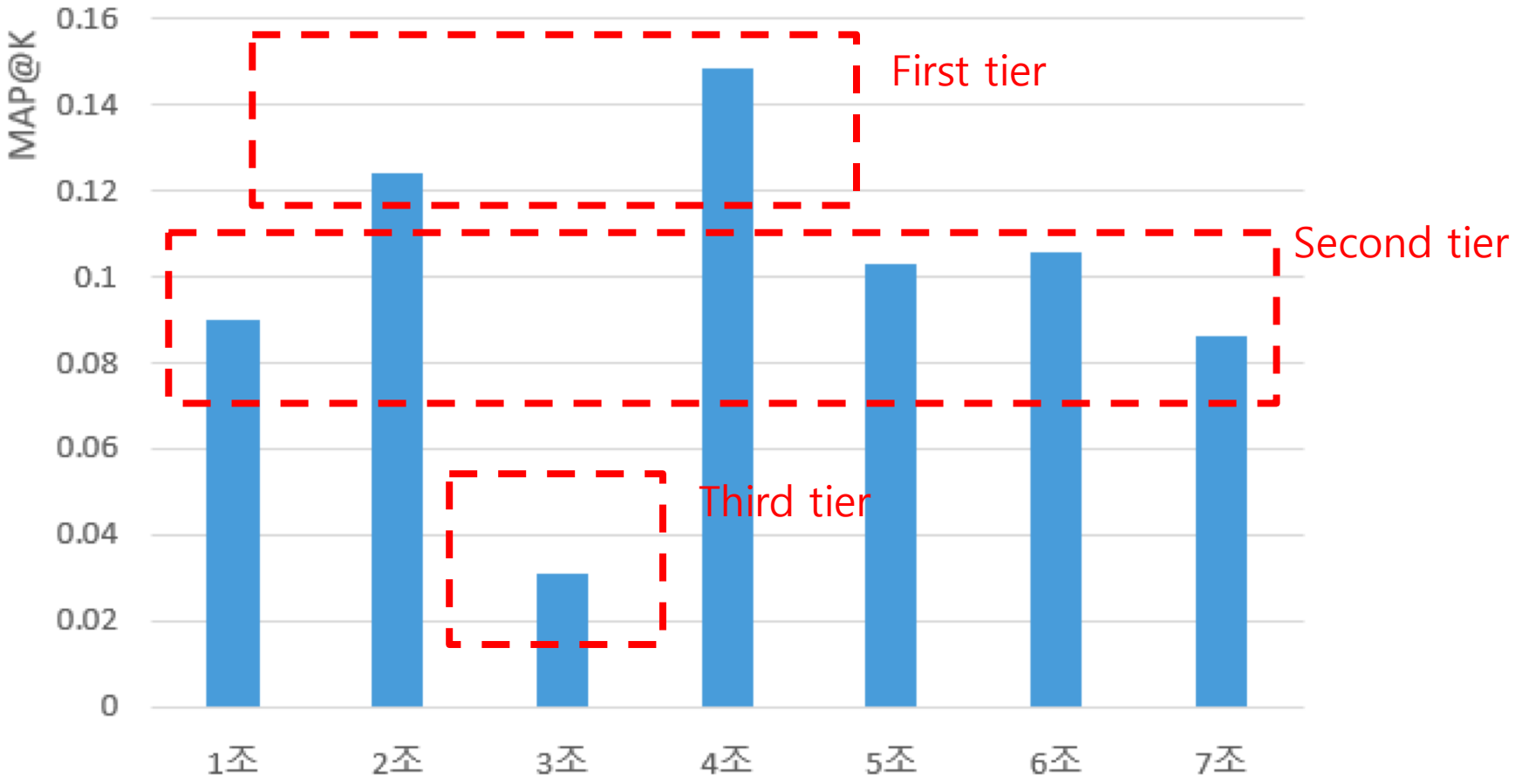


A Sample Problem

- Given
  - 1.5 years of customers behavior data from a bank
  - Contains
    - Personal attributes e.g., age and sex
    - Monthly records of products a customer has
- Goal
  - Predict what new products, such as "credit card" and "savings account", customers will purchase
- Evaluation metric
  - MAP@7

$$MAP@7 = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{min(m,7)}\sum_{k=1}^{min(n,7)}P(k)$$

Ranking of Last Year's Teams



Concerns

TAs monitor each team's attitude

- Free-rider in a team
  - Discuss the contribution of each member in the final report
  - Usually, free-riders in a team are rarely shown
- Free-rider team
  - Do not contribute any idea or information in early and final presentation
  - Condition
    - No show in early / final presentation
    - No submission
    - No contribution at all
  - Will be turned into a audience team → cannot participate competition but they can listen only
  - You can ask to be a audience

Slides

- Copyrights
  - Belong to the instructor
  - Some pages are borrowed from the Internet (without permission)
  - Slides are only for this class!
  - STRICTLY FORBIDDEN FOR RE-DISTRIBUTION
- Class github page
  - https://github.com/nongaussian/class-2018-bigdata
  - Will be switched to private repository
  - **Create github account** & join the repository next week!

Misc

- Class hours
  - Morning: Thu 10~12pm
  - Afternoon: Thu 1:00~3:00pm
  - Cancelled/back-up class: will be announced on the github page

Grading

- Ratio
  - 3 problem * 25 points (competition)
  - Final 25 points
  - Attendance -10 points
- Attendance
  - No penalty for 2 absences
    - **Includes every excuse** such as sick and 예비군
    - No need to submit any document for excuse (e.g., 결석사유서)
    - You should be in the class room in the morning and afternoon
  - From the 3th absence → -5 points
  - From the 5th absence → F
- Attitude
  - Every bad attitude will take ? points

Fix your team's table

- TA will check your team's table
- Take a table for your team in the order of arrival the next week