
Introduction to Data Science

Course introduction (Spring 2021)

Dongchul Park

dpark@sookmyung.ac.kr

Instructor and TAs

- **Instructor**

- **Dongchul Park**
- Office: Saehim 408
- Email: dpark@sookmyung.ac.kr
- Homepage: <http://cs.sookmyung.ac.kr/~dpark>
- Office hours: feel free to contact me first!

- **TAs** *274*

- Hyunju Oh (001): rophy1310@sookmyung.ac.kr
- Jiho Lee (002): jiho960214@sookmyung.ac.kr

- Special thanks to Prof. Ki Young Lee for sharing his slides for this course.

Short Bio

■ Short Bios



- **Ph.D. in CS at University of Minnesota – Twin Cities, MN, USA**
 - CRIS (Center for Research on Intelligent Storage) Lab.



- **Sr. Engineer at Samsung Semiconductor Inc., CA, USA**
 - MSL (Memory Solutions Lab)



- **Sr. Staff Engineer at Intel, OR, USA**
 - STG (Storage Technology Group)



- **Assistant Prof. at Hankuk University of Foreign Studies, Korea**
 - Computer Science & Engineering

■ Research areas

반복작성 메모리

- Big data, storage systems, SSD (NVM), key-value store, data center SW₃

Course Logistics

- **Lecture time (class room)**

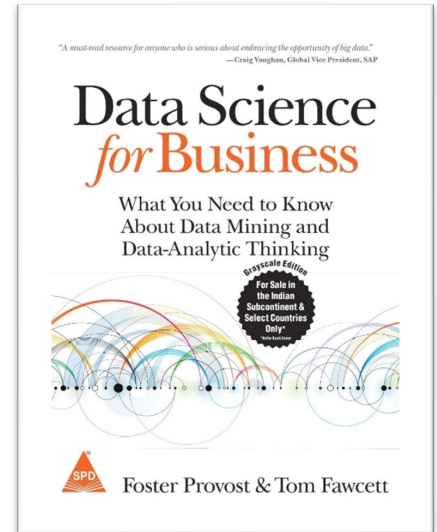
- Section 001: Tu, Th 13:30 – 14:45 (Online)
- Section 002: Tu, Th 15:00 – 16:15 (Online)

- **Course website: SnowBoard**


- Lecture slides, extra materials, notices, etc.
- **Please make sure you check this SnowBoard regularly.**

- **Materials (not limited to)**

- **Book: Data Science for Business**
 - Foster Provost & Tom Fawcett, O'Reilly, 2013.
- All slides will be posted on our website



Work and evaluation

- **Midterm: 40%**
- **Final: 40%**
- **Homework: 10%**
 - Late submission: NOT accepted. 
 - This is subject to change
- **Attendance: 10%**
 - Absence (1/4 or more): Fail with NO excuse!
 - Perfect attendance will get 10%
 - Decreases in proportion to the number of your absence
 - Elec. attendance system (+ roll call)

Contents covered

■ Main topics

- Data-analytic thinking 데이터 분석 사고
- Business problems and data science solutions
- Predictive modeling \hookrightarrow 귀납하기 \rightarrow 새로운 data의 결과 예측
- Fitting a model to data 주어진 data에 너무 잘 맞으면 X
- Avoiding overfitting 새로운 data에 유입시 문제 발생
- Similarity, neighbors, and clustering k-NN (최근접 이웃 알고리즘) 결과값 예측
- Model evaluation k-NN \rightarrow k-means
- Visualizing model performance graph화
- Evidence and probabilities
- Mining text TF-IDF age(특성) \rightarrow 22(나이값) evidence

통계 + 컴공 data mining
(분석) (데이터, 머신러닝)
 \hookrightarrow 유의미한 data의 분석 = data science

■ **Note:** Contents are subject to change based on schedules.

Questions



Dongchul Park (dpark@sookmyung.ac.kr)