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
  - 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

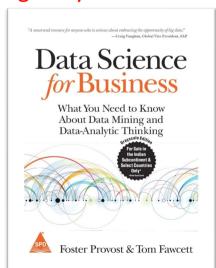- Big data, storage systems, SSD (NVM), key-value store, data center SW

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

"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

Data Science for Business

What You Need to Know About Data Mining and Data-Analytic Thinking

Grayscale Edition
For Sale in the Indian Subcontinent & Select Countries Only*

Foster Provost & Tom Fawcett

Work and evaluation

- **Midterm: 40%**
- **Final: 40%**
- **Homework: 10%**
  - Late submission: <u>NOT</u> accepted. 없는거 아십니요
  - This is subject to change

- **Attendance: 10%**
  - <u>Absence (1/4 or more): Fail with NO excuse!</u>
  - 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 분석적 사고
  - Business problems and data science solutions
  - Predictive modeling   규칙찾기 → 새로운 data의 평가예측
  - Fitting a model to data   주어진 data 기 너무잘맞으면 X
  - Avoiding overfitting   새로운 data 추후 시 안맞음
  - Similarity, neighbors, and clustering   K-NN (최고정이됐않고거리습)
    - K-NN   k-means   평균값예측
  - Model evaluation   평균%
  - Visualizing model performance   graph화
  - Evidence and probabilities   새로운 data 의예측. age(목상) → 22 (해당않나?)
  - Mining text   새로운 data 의예측.   evidence
    ex) TF-IDF

  통계 + 컴공   data mining
  (분기) (데이터, 머신러닝)
  └ 유의미한 data의 분석 = data science

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

Questions

Dongchul Park (dpark@sookmyung.ac.kr)