Principles of Data Science from Statistical Point of View

Course Code			
Class Times	Mon/Wed/Thu Type A(9:00~12:00)	**Classroom**	TBA
Equivalent Year Level	1-2	**Course Credit**	3

Instructor	Sungkyu Jung	**Sessions**	15 (45 class hours)
Office	Rm 436, Bldg 25	**Email**	sungkyu@snu.ac.kr

❑Instructor's Profile

Name: Sungkyu Jung
General Information:
Dr. Jung is an Associate Professor of Statistics at the Seoul National University. Before joining Seoul National University, he spent seven years at the University of Pittsburgh, after completing his PhD at the University of North Carolina at Chapel Hill. Dr. Jung's research interest lies in the theoretical study and applications of modern Statistics and Data Science in the analysis of data that lie on non-standard spaces. This context includes the high-dimension, low-sample-size situation, non-Euclidean data analysis, interplay between geometry and statistics, and data fusion. In particular, models and methodologies for dimension reduction, visualization of important variation and hypothesis testings need to be developed with special care for these modern data situations. Particular applications include analysis of directions, landmark-based and skeletally-modeled object shapes, data on stratified spaces or from multiple sources, and retrieving low-dimensional geometrics structures in high-dimensional data.

Education
Ph.D. in Statistics, University of North Carolina at Chapel Hill. 2011.
B.S. in Statistics, Seoul National University. 2003.

Expertise
Modern Multivariate Statistics
Non-Euclidean data analysis
Data fusion

Most Recent Works

- Sungkyu Jung, Myung Hee Lee and Jeongyoun Ahn (2018). "On the number of principal components in high dimensions," *Biometrika* 105(2), 389-402.
- Sungkyu Jung (2018). "Continuum directions for supervised dimension reduction," *Computational Statistics and Data Analysis* 125, 27-43
- Gen Li and Sungkyu Jung (2017). "Incorporating Covariates into Integrated Factor Analysis of Multi-View Data," *Biometrics* 73 (4), 1433-1442.
- David Groisser, Sungkyu Jung, and Armin Schwartzman (2017). "Geometric foundations for statistics on symmetric positive definite matrices: characterizations of minimal scaling-rotation curves in low dimensions", *Electronic Journal of Statistics*, Vol. 11, No. 1, 1092-1159.

❑Course Information

Course Description	Data science is an emerging interdisciplinary field stemming from statistics, mathematics and computer science. At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. The objective of this course is to provide students with a principled introduction to data science that properly combines inferential thinking and computational thinking. Students will learn the fundamental pipeline of data science, ranging from data acquisition, data clean-up, data exploration and visualization, modeling and inference, to professional reporting. Students will learn data science by doing data science via statistical language R using RStudio. Students will get their hands "dirty" by manipulating and analyzing real data from politics, economics, industry, medical science and sports. No prior experience on programming or introductory statistics is required.
Course Evaluation	Homework, class project, class participation 40% Midterm exam 30% Final exam 30% **Attendance** will be important for keeping up with class. Good attendance and **active participation** will be reflected in grade.
Course Materials	Textbook: Baumer et al., 2017. Modern Data Science with R. CRC Press. ISBN 9781498724487 Lecture notes and slides, reading assignments will be available e-TL (http://etl.snu.ac.kr), SNU's online course management system.
Class Policy	Academic integrity: Students have an obligation to exhibit honesty and to respect the ethical standards of the academy in carrying out his or her assignments.

❑ Course Schedule

Session 1: Introduction to Data Science
- Course overview
- Introduction to Data Science and Data Science tools
- RStudio interface, R basics, packages and Rmarkdown

Session 2: Data Visualization
- Statistical graph: Visualizing distributions of numbers and categories
- Tufte's design principles

Session 3: Dissecting Data Graphics
- Four elements of data graphics
- Grammar of Graphics and ggplot

Session 4: Data Visualization tools
- Statistical transformations
- 3D and interactive graphics

Session 5: Data Wranging - one table
- Five verbs of data manipulation
- Tibble and pipe operator

Session 6: Data Wrangling – two tables
- Relational data manipulation

Session 7: Tidy data
- Data cleaning
- Data preparation

Session 8: Mid-term exam

Session 9: Statistical Foundation
- Exploratory data analysis
- Samples and populations
- Data summary and uncertainty quantification

Session 10: Sampling distributions
- Confidence interval and hypothesis testing
- Sampling distribution by resampling

Session 11: Modeling associations
- Conditional modeling
- Correlation
- Regression

Session 12: Cause and effect
- Causality
- Confounding effect and
- Randomization

Session 13: Text Mining
- Handling text data
- Word cloud
- Sentiment analysis

Session 14: Professional ethics and outlook for handling Big Data
- Principles to guide ethical action
- Reproducible research
- Garden of forking paths
- Outlook for Big Data analysis

Session 15: Final exam