Sungkyunkwan University (SKKU) International Summer Semester (ISS) 2018

Data Science and Management

Prof. Dr. David T. Reitter, The Pennsylvania State University

SHORT COURSE DESCRIPTION

When the English physician John Snow discovered the source of a cholera epidemic, he also discovered new scientific methodology: inference from multi-layered data. Snow counted cholera cases and plotted them on a map, leading to the neighborhood's water pump shared by the patients. Granted: no science is possible without data. What has changed? Massive-scale data collections have become available, as have new statistical and computational methodologies. Since the 1990's, the use of big data and machine learning techniques has revolutionized Artificial Intelligence, leading to computer programs that understand natural language, answer questions and drive cars. Data in many new areas allow us to ask many new questions, from smart cities and energy sustainability to disease control and medicine.

Inference from big data has the potential to change science, but that is controversial. The methods paint a superficial, easily biased picture of the data that needs to be interpreted to form a true understanding of the causal structure of the world around us.

This course teaches practical exploratory data analysis, statistical modeling as well as the associated caveats. Upon completing the course, students will be able to ask focused questions about datasets, visualize data, and critically evaluate empirical claims.

Pre-Requisites: This course will involve a fair amount of practical programming and statistical analysis. Students need to be comfortable with a practically useful programming language or with math (typically, two college-level courses and some experience with programming, or skills in abstraction and mathematical thinking). **Do not take this class if you are not comfortable with programming or math.** Knowing how to use a Unix command line (grep, cat) is helpful. The course will provide an introduction to Python, and a light-weight introduction to using R. However, but it will not explain basic programming knowledge (such as about control structures or data structures). The course will introduce statistical modeling, but ideally, participants should have taken an introductory probability theory course (probabilities, distributions, hypothesis testing).

Each participant needs to bring their own laptop.

ABOUT YOUR INSTRUCTOR

Professor **David Reitter** (reitter@psu.edu) is your instructor for this course. He is an Assistant Professor of Information Science and Technology at Penn State, holds a PhD from the University of Edinburgh, and was a post-doctoral fellow at Carnegie Mellon University. His research helps us understand how the mind produces and comprehends natural language, and how we can use this knowledge to revolutionize artificial intelligence. Reitter co-directs the Applied Cognitive Science lab at Penn State. Dr. Reitter's work is funded by the USA's National Science Foundation. He is also the creator of the widely popular *Aquamacs* software.

READING MATERIALS

The lectures provide the core content of the course, introducing key theories and research findings. The information is supplemented by readings from the textbook, and by other articles included in the homework assignments. Articles associated with the homework assignments will be available to download from the website. Not all material in the readings will be covered in the lecture, and vice versa, so it is important to keep up with both.

COURSE REQUIREMENTS AND GRADING

After each class, you need to fill out a "log entry". This is your dairy for this semester. You may reflect on what was discussed in the lecture, relate it to personal experience or research results you are aware of, and you may provide feedback on what you liked or disliked in class. There may be a mid-term and a final exam, which will contain a mix of multiple-choice and essay questions. All answers must be given in English. The final grade is figured as 20% mini-exam and graded exercise results, 40% critical reflection/summary, 30% in-class presentation (if given), 10% participation. This weighting is subject to adjustment. Grounds for failing the class include failure to reach at least 60% in grades; failure to show up for most classes, or for the exams, and academic dishonesty.

COURSE SCHEDULE

Most classes consist of a lecture followed by a practical application of the concepts learned. Students will, as individuals and in groups, concrete prepare, process, and analyze data. A challenge serves as a capstone project for the course; students will be given a dataset to analyze and visualize (e.g., citation networks: Standard Large Network Dataset Collection, health & longevity: The Human Mortality Database, finance: geographical stock price correlations, Quandl datasets; Global Warming / Sustainability: Berkeley Earth Data). Students will take a new dataset, develop ideas and theories before exploring the data, testing their hypotheses and communicating the results as a research poster.

– WEEK I –

Crash Course: practical data processing

- Practical: Python, pandas, scipy and their installation – inference from big data
- Practical: Programming: collecting and preparing data

Statistical inference: from small to big data

- Graphing. Visual data exploration. How to plot data.

– WEEK II –

- Practical (in teams): analyze and visualize a dataset.
- Creative (in teams): Make an infographic from the dataset.
- Manipulative, controlled experiments: finding out if A causes B
- Hypothesis testing and its caveats.
- Correlation is not causation: Why we need to build causal models to understand the world, and why correlational inference is not enough. On the value of experiments in an observational world
- Linear Regression Modeling, structured random effects
- Technical solutions to data management and parallel computation: NoSQL databases, MapReduce, Hadoop. Crash-course in the SQL database query language.

– WEEK III –

The human dimension

- Human illusions. How to lie with data.
- The human dimension: Cognitive Biases that lead us to see what we want to see, and to never try to disprove our own theories

Advanced modeling

- Network Science (overview only): An introduction to graph-theoretic analysis of sociotechnical relationships. PageRank, Google's innovation that changed search on the web.
- Complex models (overview only): Machine Learning, e.g., Categorization with Deep Neural Networks. Bayesian Modeling.

– WEEK IV –

- Challenge: Analyze a big dataset