Introduction to Big Data and Machine Learning
SURV XXX

Term

Instructor(s):

**Trent D. Buskirk**, PhD, is Vice President of Statistics and Methodology at Marketing Systems Group. Formerly a Research Director for the Nielsen Company and Associate Professor of Biostatistics in the Department of Biostatistics of the Saint Louis University School of Public Health, Dr. Buskirk has been conducting research relating to the use of cell phones and smartphones in survey research for over 15 years. His research interests also include dual frame weighting for cell phone surveys, as well as mode effects related to cell phone surveys, online and in-person surveys. Dr. Buskirk's research work has appeared in various journals including the *Journal of Official Statistics, Journal of the Royal Statistical Society, Field Methods, Social Science Computer Review and Survey Practice*. He is currently the Chair of the American Statistical Association's Survey Review Committee, Publications Officer for ASA's Survey Research Section and an active member of AAPOR's Address Based Sampling and Future of Telephone Surveys Task Forces.

**Frauke Kreuter**, PhD, is Professor at the Joint Program in Survey Methodology and Professor of Statistics and Social Science Research Methods at the University of Mannheim. Before joining the University of Maryland she held a postdoc at the UCLA Statistics Department. Dr. Kreuter is Fellow of the *American Statistical Association*, and currently Associate Editor of the *American Sociological Review, Journal of Official Statistics, Journal of Statistical Software, Sociological Methods and Research, and Survey Research Methods*. Dr. Kreuter has been conducting research related to measurement error and nonresponse in social surveys for roughly twenty years. She is known among the AAPOR community for her work on paradata, to which she has taught numerous short courses in the past. Currently Dr. Kreuter is co-chairing the AAPOR Task Force on Big Data.

fkreuter@umd.edu

Class Structure and Course Concept:

This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video recorded lectures and reading the required literature for each unit and then “attending” mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor. There will be quickly homework and quizzes, as well as a final exam.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, for each credit offered by a course, students can expect to spend one hour per week on in-class activities and three hours per week on out-of-class activities over the span of a full 12-week term. This is a 1-credit course that runs for 4 weeks. Hence, the total average workload is about 12 hours per week.

Mandatory Weekly Online Meetings:

Tuesday 6pm EST
Meetings will be held online through BlueJeans. Follow the link to the meeting sessions on the course website on jpsmonline.umd.edu. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to e-mail questions about the materials covered in the videos and readings of the week to the instructor (fkreuter@umd.edu) before the meetings (deadline for sending questions via e-mail is Monday 6pm).

Overview of the Course
The amount of data generated as a by-product in society is growing fast including data from satellites, sensors, transactions, social media and smartphones, just to name a few. Such data are often referred to as "big data", and can be used to create value in different areas such as health and crime prevention, commerce and fraud detection. Big Data are often used for prediction and classification tasks. Both of which can be tackled with machine learning techniques. In this course we explore how Big Data concepts, processes and methods can be used within the context of Survey Research. Throughout this course we will illustrate key concepts using specific survey research examples including tailored survey designs and nonresponse adjustments and evaluation. This course will offer participants:

- an overview of key Big Data terminology and concepts
- an introduction to common data generating processes
- a discussion of some primary issues with linking Big Data with Survey Data
- issues of coverage and measurement errors within the Big Data context
- a discussion of information extraction and signal detection in the context of Big Data
- a discussion of the similarities and differences in model building for inference versus prediction
- an overview of general concepts from machine learning as they apply to processing Big Data
- a discussion of signal detection and information extraction
- a discussion of the potential pitfalls for inference from Big Data
- an introduction to a small set of key analytic techniques (e.g. classification trees, random forests, conditional forests) to process Big Data using R with example code provided

Grading
Grading will be based on:

- 4 online quizzes (worth 10% total)
- Participation in discussion during the weekly online meetings and submission of questions via e-mail (deadline: Monday, 8AM before class) demonstrating understanding of the required readings and video lectures (30% of grade)
- Project (40%) consisting of 4 homework assignments (worth 10% each)
- A final open-book online exam (20% of grade)
Dates of when assignment will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructors.

**Technical Equipment Needs:**
The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a standalone Web camera and a headset. Decent quality headsets and webcams are available for less than $20 each. We ask students to refrain from using built-in Web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

**Resources and Readings**

http://www.dataminingconsultant.com/resources.htm


An Introduction to Statistical Learning with Applications in R (2013) Free PDF Version: [http://bit.ly/1iUJso0](http://bit.ly/1iUJso0)

Online Resources for FREE lecture videos and labs in R [http://bit.ly/1snBMk5](http://bit.ly/1snBMk5)
An overview of Machine Learning Functions available in R [http://cran.r-project.org/web/views/MachineLearning.html](http://cran.r-project.org/web/views/MachineLearning.html)

Additional required and recommended readings will be made available on the course website: jpsmonline.umd.edu

Articles and books for course examples


**Academic Conduct**

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at the University of Maryland Graduate School web site

[http://www.graduate.umaryland.edu/policies/misconduct.html](http://www.graduate.umaryland.edu/policies/misconduct.html)

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

**Accommodations for Students with Disabilities**

In order to receive services you must contact the Disability Support Services (DSS) office to register in person for services. Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301.314.7682; [http://www.counseling.umd.edu/DSS/](http://www.counseling.umd.edu/DSS/).

**Course Evaluation**

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in a course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

**Class Schedule**

Unit 1: Overview of Big Data
*What is Big Data? Why does it matter? Why is this relevant for survey researchers?*

Unit 2: Working with Big Data

Unit 3: Classical Statistical Approaches versus Statistical Machine Learning

Unit 4: Model Evaluation/Validation

Unit 5: K-Means Clustering + K Nearest Neighbors

Unit 6: CARTS + Random Forests