Spring 2021
Big Data and Machine Learning for IO Psychology (PSYC 592/ PSYC892)
Tuesday and Thursday 3:00 PM-4:15 PM
Online

PROFESSOR: Philseok Lee, Ph.D.
ZOOM ONLINE OFFICE HOURS: by appointment
OFFICE: 3056 David King Hall
EMAIL: plee27@gmu.edu

LECTURES: T, TR: 3:00-4:15 pm online

TEXTBOOK: There is NO required textbook for this course. The course materials are made from various textbooks as well as recent research literature. However, here are some recommended books and blogs.

USEFUL BOOKS OR BLOGS:

(1) Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media. This book can be freely downloaded from https://web.stanford.edu/~hastie/ElemStatLearn//printings/ESLII_print12_toc.pdf
(2) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning, New York: springer. This book can be freely downloaded from http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf
(3) García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (pp. 195-243). Cham, Switzerland: Springer International Publishing. This book can be freel downloaded from https://link.springer.com/content/pdf/10.1007/978-3-319-10247-4.pdf
(4) Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer.
(5) Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.
(6) Ledolter, J. (2013). Data mining and business analytics with R. John Wiley & Sons.
(7) Tonidandel, S., King, E. B., & Cortina, J. M. (Eds.). (2015). Big data at work: The data science revolution and organizational psychology. Routledge.
(8) Woo, S. E., Tay, L., & Proctor, R. W. (Eds) (2020). Big data in psychological research. American Psychological Association
(9) What's the Big Data (https://whatsthebigdata.com/): Curates a lot of quick, interesting things on data science in general, largely industry focused
(10) Better Explained (https://betterexplained.com/): A good resource for accessible math explainers
(11) The R Graph Gallery https://www.r-graph-gallery.com/
(12) R Graphics Cookbook, https://r-graphics.org/
(13) Toward Data Science, https://towardsdatascience.com/

COURSE DESCRIPTION:

In the current workplace, big data are more relevant and accessible than ever (e.g., data from surveys, sensors, social media, smartphones, and documents, just to name a few). Big data describe the amount and complexity of the data, as well as the analytical technique and tools revolving around it. As the volume, velocity, and variety of organizational data increase, IO reearchers and practioners have to adapt to nontraditional approaches that harness the data at an unprecedented level. Further, the development of statistical, computational, and data management methods in Machine Learning allows us to gain additional insights by applying novel methods to investigate familiar topics with "big" and "small" data. Big Data are often used for prediction and classification tasks. Both of which can be tackled with machine learning techniques. This course will be a generic introduction to the concept and application of Big Data in IO psychology and how to analyze the data using Machine Learning techniques.
During this course you will:

- Identify practical problems which can be solved with machine learning
- Learn the R statistical programming language via building simple data visualizations
- Learn how to perform data wrangling, tidying, and visualization using packages from the dplyr and ggplot2
- Understand mathematical and statistical concepts of various Machine Learning models
- Build, tune and apply various Machine Learning models with Rstudio (e.g., decision-tree, naïve bayes, various regression models (logistic, lasso, lidge, elastic net, and nonlinear regressions), association rule mining, support vector machine, artificial neural network, ensemble methods (e.g., random forests, boosting, bootsrap aggregation, gradient boosting, and stacked ensemble), and text mining.

SOFTWARE:

R is chosen as the programing language for this course, recognizing the growing importance of R programing in data science as well as its great utility in modeling. This course will dedicate the first two weeks of lectures for R programing, which is the programing language used for instruction and student projects. Topics will include three parts:

- The first is on the basic programing language features. This includes data structures such as vectors, lists, arrays, matrices, data frames; structured programing constructs such as loops, conditional statements and functions; data manipulation tools, file I/Os, etc.
- The second is on R functions for graphics and visualization.
- The third is on the statistical aspect of R, which covers various R functions for statistical tests.

Sample R codes will be provided for most of the examples, so that you can try R programming on your own and gain hands-on experience. You can download R (https://www.r-project.org/) and R studio (https://www.rstudio.com/products/rstudio/download/).

Here are some useful materials for R:

- An Introduction to R by Venables et al. (2013) https://cran.r-project.org/doc/manuals/R-intro.pdf
- Quick-R https://www.statmethods.net/
- Cookbook for R http://www.cookbook-r.com/

COURSE STRUCTURE, GRADING, AND REQUIREMENTS:

1. Attendance and Participation:

It will count 5% toward your final grade.

- Given this is a virtual class, I will check your attendance. You will lose 0.17% for each day you miss the class.
- My expectation is that you will attend every class meeting and engage during class sessions.

2. Homework:

The homework assignments will count 50% toward your final grade. There will be 7 homework assignments.

- Please submit your assignments through Blackboard by noon on the due date. Assignments received after 12 pm will be late.
- You will lose 10% for each day the assignment is late (with the first day beginning after 12 pm the day it was due).
- Homework submitted more than a week late will not be accepted and a grade of zero will be given.
- You are free to discuss the assignments with other students in the class. However, you MUST complete the work INDEPENDENTLY.

3. Final Project:

Group project will count 35% toward your final grade: Final paper and submitted materials will count 30% & class presentation will count 5%. Thus, total would be 35%.
The course culminates with a final project, completed throughout the semester in small group. Students must craft a well-defined research question and then identify and analyze real datasets to answer that question. Here are detailed information about final project.

Project Overview

Final Project will give you the chance to explore a topic of your choice and to expand your analytical skills. By working with real data of your choosing you can examine questions of particular interest to you.

The broad objectives for the project are to:

- Identify the problems and goals of a real situation and dataset.
- Choose an appropriate approach for formalizing and testing the problems and goals, and be able to articulate the reasoning for that selection.
- Implement your analysis choices on the dataset.
- Interpret the results of the analyses.
- Contextualize those results within a greater scientific and social context, acknowledging and addressing any potential issues related to privacy and ethics.
- Work effectively to manage a project as part of a team.

To accomplish this you will work in teams of **2 students** to conceive of and carry out an analysis project.

Guidelines for Project Proposal

The Project Proposal should be 1-2 page (single space, 12 pts). The Proposal template can be found in **Appendix 1**. Your proposal must include the following sections:

- NAMES: Be sure to include each member's name
- RESEARCH QUESTION: What is your research question? Include the specific question you're setting out to answer. This question should be specific, answerable with data, and clear.
- BACKGROUND AND/OR PRIOR WORK: Present the background and context of your topic (i.e., a brief literature review) and question in a few paragraphs. Include a general introduction to your topic and then describe what information you currently know about the topic after doing your initial research. Include references to other projects who have asked similar questions or approached similar problems. Explain what others have learned in their projects.
- DATA: Describe the datasets you would like to explore.
    - What is the source?
    - What variables you have?
    - How they were collected?
    - How many observations you have?
    - What/who are the observations? Over what time period? etc.
- TEAM EXPECTATIONS: Include your group's expectations of one another for successful completion of your project. Discuss how your team will communicate throughout the semester.
- PROJECT TIMELINE PROPOSAL: Specify your team's specific project timeline. An example timeline has been provided. Changes the dates, times, names, and details to fit your group's plan.

Project Submission

- Students should choose their interested reseach topic and found any open source dataset. Build, tune and apply various Machine Learning models. **You should submit final paper by May 4th noon. You should follow the SIOP Poster Session format (A summary with a maximum of 3,000 words.** But, references, tables, and figures do not count toward the limit). You can find a detailed information from https://www.siop.org/Annual-Conference/Registration-and-Resources/Call-for-Proposals/Preparing-and-Formatting-Your-Proposal-Document
- **Students will have to make presentation in class (15-20 minutes) for the last week of this class .**
- The project submission includes all the code used for all components of thre project, as well as written final paper and data visualizations.
- **Final paper and submitted materials will count 30% & class presentation will count 5%. Thus, total would be 35%.**
- Check Points
    - ABSTRACT: Summarizing your group's project and results.
    - NAMES: See proposal specifications.
    - RESEARCH QUESTION: See proposal specifications.
    - BACKGROUND & PRIOR WORK: See proposal specifications.

- DATASET(S): Describe your dataset
- DATA CLEANING: Describe your process
- DATA ANALYSIS & RESULTS: For examples:
  - What distributions do your variables take?
  - Are there any outliers?
  - Relationship between variables?
  - Analysis (Note that you will likely have to do some Googling for analytical approaches not discussed in class. This is expected for this project and an important skill for a data scientist to master.)
  - What approaches did you use? Why?
  - What were the results?
  - What were your interpretation of these findings.

- DATA VISUALIZATION - There must be at least three (3) appropriate data visualizations throughout these sections. Each visualization must included an interpretation of what is displayed and what should be learned from that visualization. Be sure that the appropriate type of visualization is generated given the data that you have, axes are all labeled, and the visualizations clearly communicate the point you're trying to make.
- ETHICS & PRIVACY: Be sure to update with what you actually did to take the ethical considerations into account for the analysis you did.
- CONCLUSION & DISCUSSION: Discuss your project. Summarize your data and question. Briefly describe your analysis. Summarize your results and conclusions. Be sure to mention any limitations of your project. Discuss the impact of this work.

How to Find Datasets ?

The purpose of this project is to find a real-world problem and dataset (or likely, datasets!) that can be analyzed with the techniques learned in class and those you learn on your own. **The best datasets are the ones that can help you answer your question of interest.** Here are example sources to find datasets, but you are welcome to use other sources.

https://www.apa.org/research/responsible/data-links
https://www.forbes.com/sites/bernardmarr/2018/02/26/big-data-and-ai-30-amazing-and-free-public-data-sources-for-2018/?sh=65b8fafd5f8a
https://www.data.gov/
https://data.gov.uk/
https://ieee-dataport.org/datasets
https://dataverse.harvard.edu/
https://github.com/awesomedata/awesome-public-datasets
http://archive.ics.uci.edu/ml/index.php
https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit?usp=sharing
https://ucsd.libguides.com/data-statistics/home
https://data.stanford.edu/
https://www.census.gov/
https://dasl.datadescription.com/datafiles/

Or you can find dataset from other sources.

4. Quiz:

Quizzes will count 10% toward your final grade.

- There will be 10 quizzes and each quiz will count 1% of your final grade.
- Quiz will be simple concepts or applications of big data and machine learning techniques (short answer or multiple choice questions).
- This should be an INDEPENDENT work. You SHOULD NOT discuss with other students.

Grading:

Letter grades will be determined according to the following scheme:

A+ 98-100%	A 93-97.9%	A- 90-92.9%	B+ 87-89.9%
B 83-86.9%	B- 80-82.9%	C+ 77-79.9%	C 73-76.9%
C- 70-72.9%	D+ 67-69.9%	D 63-66.9%	D- 60-62.9%
F Below 60%			

Failure to complete course requirements:

Students who miss a small portion of the course due to an excused absence may be given a grade of Incomplete (I). However, failure to complete the required work by the end of the following semester will result in a grade of F.

Missing class due to religious observances:

Students who anticipate the necessity of being absent from class due to the observation of a major religious observance must provide notice of the date(s) to the instructor, in writing.

ACADEMIC INTEGRITY

Homework: You can discuss and work on assignments with other students, but each student must turn in his/ her own results and interpretation. For example, when analyzing data, students might discuss how to approach a problem, run their own analyses, and independently write up the results. Simply copying someone else's answers verbatim is unacceptable and will be subject to loss of points or a zero credit.

Quiz: Students must work alone on each quiz. If you have questions about problems or potential solutions, consult with your instructor only.

Failure to follow these guidelines may be viewed as evidence of academic dishonesty, which can result in a grade of FF for the course and other penalties through the University System.

DISABILITY SERVICES

Disability Services at George Mason University is committed to providing equitable access to learning opportunities for all students by upholding the laws that ensure equal treatment of people with disabilities. If you are seeking accommodations for this class, please first visit

http://ds.gmu.edu/ for detailed information about the Disability Services registration process. Then please discuss your approved accommodations with me. Disability Services is located in Student Union Building I (SUB I), Suite 2500. Email: ods@gmu.edu | Phone: (703) 993-2474

IMPORTANT DATES

Please check Fall 2021 – Drop / Withdrawal Deadline Changes from this link
https://registrar.gmu.edu/calendars/spring-2021/

Last day to add a class: February 1
Last day to drop (with 100% tuition refund): February 16
Selective withdrawal period: March 2 – April 1
Last day of classes: April 30
Reading day(s): May 1
Final exam period: May 3 –May 10

COURSE SCHEDULE, READINGS, AND ASSIGNMENTS

*Note: While we certainly will try to adhere to this schedule, we may need to rearrange things a bit during the semester.

Hastie et al.(2009) indicates "Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media." This book can be downloaded from https://web.stanford.edu/~hastie/ElemStatLearn//printings/ESLII_print12_toc.pdf

* ***James et al.(2013)*** indicates "James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning, New York: springer". This book can be downloaded from http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

*** *García et al.(2015)*** indicates García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (pp. 195-243). Cham, Switzerland: Springer International Publishing. This book can be downloaded from https://link.springer.com/content/pdf/10.1007/978-3-319-10247-4.pdf

Week	Date	Topic	Recommended Book Chapters and Resources	Assignment	Quiz
1	26-Jan	Intro to Big Data and Machine Learning for IO Psychology	Boyd & Crawford (2012). Chapter 2 (Introduction to statistical learning) from Hastie et al. (2009)		
	28-Jan	Introduction to R & Visualization	Culpepper, S. A., & Aguinis, H. (2011) https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf http://www.cookbook-r.com/ https://r-graphics.org/		
2	2-Feb	Introduction to R & Visualization	https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf http://www.cookbook-r.com/ https://r-graphics.org/	**Assignment 1 (Intro to R & Visualization: 10%) OUT**	Quiz1 (1%)
	4-Feb	Data Wrangling with dplyr	https://ohi-science.org/data-science-training/dplyr.html https://moderndive.com/3-wrangling.html		

3	9-Feb	Data Wrangling with dplyr	https://ohi-science.org/data-science-training/dplyr.html https://moderndive.com/3-wrangling.html	Assignment 1 (10%) DUE	
	11-Feb	Data Preprocessing for Machine Learning	Chapter 3, 4, 5 from García et al. (2015)		
4	16-Feb	Data Preprocessing for Machine Learning	Chapter 3, 4, 5 from García et al. (2015)	**Assignment 2 (Data Management & Preprocessing: 10%) OUT**	Quiz2 (1%)
	18-Feb	Decision Tree & Regression Tree Models	Chapter 5 & 8 from James et al. (2013) Chapter 7 & 9 from Hastie et al. (2009)		
5	23-Feb	Decision Tree & Regression Tree Models	Chapter 5 & 8 from James et al. (2013) Chapter 7 & 9 from Hastie et al. (2009)	Assignment 2 (10%) DUE	
	25-Feb	Decision Tree & Regression Tree Models	Chapter 5 & 8 from James et al. (2013) Chapter 7 & 9 from Hastie et al. (2009)	**Project Proposal DUE**	Quiz3 (1%)
6	2-Mar	Naïve Bayes	TBA		
	4-Mar	Naïve Bayes	TBA	**Assignment 3 (Decision Tree Models & Naïve Bayes: 10%) OUT**	Quiz4 (1%)
7	9-Mar	LASSO, Lidge, Elastic Net, Nonlinear Regression	Chapter 3 from Hastie et al. (2009) Chapter 6 & 7 from James et al. (2013)		
	11-Mar	LASSO, Lidge, Elastic Net, Nonlinear Regression	Chapter 3 from Hastie et al. (2009) Chapter 6 & 7 from James et al. (2013)	Assignment 3 (10%) DUE	Quiz5 (1%)
8	16-Mar	Association Rule Mining	Chapter 14.1 from Hastie et al. (2009)	**Assignment 4 (Regression Models: 10%) OUT**	
	18-Mar	Association Rule Mining	Chapter 14.1 from Hastie et al. (2009)	**Assignment 4 (Regression Models and Association Rule: 10%) OUT**	Quiz6 (1%)

9	23-Mar	Support Vector Machine	Chapter 9 from James et al. (2013) Chapter 12 from Hastie et al. (2009)		
	25-Mar	Support Vector Machine	Chapter 9 from James et al. (2013) Chapter 12 from Hastie et al. (2009)	Assignment 4 (10%) DUE	
10	30-Mar	Support Vector Machine	Chapter 9 from James et al. (2013) Chapter 12 from Hastie et al. (2009)	**Assignment 5 (Support Vector Machine: 10%) OUT**	Quiz7 (1%)
	1-Apr	Artificial Neural Network (ANN)	Chapter 11 from Hastie et al. (2009)		
11	6-Apr	Artificial Neural Network (ANN)	Chapter 11 from Hastie et al. (2009)	Assignment 5 (10%) DUE	
	8-Apr	Artificial Neural Network (ANN)	Chapter 11 from Hastie et al. (2009)	**Assignment 6 (Artificial Neural Network: 10%) OUT**	Quiz8 (1%)
12	13-Apr	Ensemble method: Bootstrap Aggregation, Random Forest, Boosting, Gradient Boosting & Stacked Ensemble	Chapter 10, 15, 16 from Hastie et al. (2009)		
	15-Apr	Ensemble method: Bootstrap Aggregation, Random Forest, Boosting, Gradient Boosting & Stacked Ensemble	Chapter 10, 15, 16 from Hastie et al. (2009)	Assignment 6 (10%) DUE	
13	20-Apr	Ensemble method: Bootstrap Aggregation, Random Forest, Boosting, Gradient Boosting & Stacked Ensemble	Chapter 10, 15, 16 from Hastie et al. (2009)	**Assignment 7 (Ensemble method: 10%) OUT**	Quiz9 (1%)
	22-Apr	Intro to Text Mining (Webscraping, Topic Modeling, Sentiment Analysis)	Pennebaker, Mehl, & Niederhoffer (2003). Kjell, O. N., Kjell, K., Garcia, D., & Sikström, S. (2019). Schmiedel, T., Müller, O., & vom Brocke, J. (2019).		

			Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2018). Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018).		
14	27-Apr	Intro to Text Mining (Webscraping, Topic Modeling, Sentiment Analysis)	Pennebaker, Mehl, & Niederhoffer (2003). Kjell, O. N., Kjell, K., Garcia, D., & Sikström, S. (2019). Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2018). Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018).	Assignment 7 (10%) DUE	
	29-Apr	Intro to Text Mining (Webscraping, Topic Modeling, Sentiment Analysis)	Pennebaker, Mehl, & Niederhoffer (2003). Kjell, O. N., Kjell, K., Garcia, D., & Sikström, S. (2019). Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2018). Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018).		Quiz10 (1%)
15	4-May	Class Presentation (5%)		**Project Paper Due (25%)**	
	6-May	Class Presentation (5%)			

Note: All topics and dates are subject to change. Schedule may be revised as the semester proceed.

Recommended I-O Related Reading Lists:

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 505-533.

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, *34*, 826-844.

Alexander III, L., Mulfinger, E., & Oswald, F. L. (2020). Using Big Data and Machine Learning in Personality Measurement: Opportunities and Challenges. *European Journal of Personality*, *34*, 632-648.

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020b). Machine learning in psychometrics and psychological research. *Front. Psychol.* 10:2970. doi: 10.3389/fpsyg.2019.02970

Fokkema, M., & Strobl, C. (2020). Fitting prediction rule ensembles to psychological research data: An introduction and tutorial. *Psychological Methods*.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*, 809-816.

Lee, A., Inceoglu, I., Hauser, O., & Greene, M. (2020). Determining causal relationships in leadership research using Machine Learning: The powerful synergy of experiments and data science. *The Leadership Quarterly*, 1- 14

Flesia, L., Monaro, M., Mazza, C., Fietta, V., Colicino, E., Segatto, B., & Roma, P. (2020). Predicting Perceived Stress Related to the Covid-19 Outbreak through Stable Psychological Traits and Machine Learning Models. *Journal of clinical medicine*, *9*(10), 3350.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*, 190-203.

Mazza, C., Monaro, M., Orrù, G., Burla, F., Colasanti, M., Ferracuti, S., & Roma, P. (2019). Introducing machine learning to detect personality faking-good: a new model based on MMPI-2-RF scales and reaction times. *Frontiers in Psychiatry*, *10*, 389-406.

Kjell, O. N., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*, 92.

Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, *22*, 941-968.

Stachl, Clemens, Quay Au, Ramona Schoedel, Daniel Buschek, Sarah Völkel, Tobias Schuwerk, Michelle Oldemeier et al. "Behavioral patterns in smartphone usage predict big five personality traits." (2019).

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). "Where's the IO?" Artificial IntellIgence and Machine Learning in talent management systems. *Personnel Assessment and Decisions*, *3*, 33-44.

Hamilton, I. A. (2018, October, 10). Amazon built AI to hire people, but it discriminated against women. *Business Insider,* Retrived from https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10

Lavanchy, M. (2018). Amazon's sexist hiring algorithm could still be better than a human. The Conversation.

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society, 23,* 128-147.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods, 21*, 525-547.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning* (pp. 197-253). Springer, Cham.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2018). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 1-33.

Tay, L., Ng, V., Malik, A., Zhang, J., Chae, J., Ebert, D. S., ... & Kern, M. (2018). Big data visualizations in organizational science. *Organizational Research Methods*, *21*(3), 660-688.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, *14*, 91-118.

Campion, M. C., Campion, M. A., & Campion, E. D. (2018). Big data techniques and talent management: Recommendations for organizations and a research agenda for IO Psychologists. *Industrial and Organizational Psychology*, *11*, 250-257.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods*, *21*, 733-765.

Ihsan, Z., & Furnham, A. (2018). The new technologies in personality assessment: A review. *Consulting Psychology Journal: Practice and Research*, *70*, 147-166.

Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: theoretical and methodological issues. *Journal of Computational Social Science*, *1*, 59-66.

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*, 246-257.

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*, 899-917.

Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media and Society, 4*, 2056305118768300.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*, 689-732.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, *43*, 5-18.

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior research methods*, *49*, 1630-1638.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100-1122.

Bleidorn, W., Hopwood, C. J., & Wright, A. G. (2017). Using big data to advance personality theory. *Current opinion in behavioral sciences*, *18*, 79-82.

Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. Frontiers in psychology, 7, 738.

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological methods*, *21*, 603-620.

Oswald, F. L., & Putka, D. J. (2015). Statistical methods for big data: A scenic tour. In *Big Data at Work* (pp. 57-77). Routledge.

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological methods*, *21*, 493-506.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, *21*, 1-19.

Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social media in employee-selection-related decisions: A research agenda for uncharted territory. *Journal of management*, *42*, 269-298.

Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological methods*, *21*, 583-602.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*, 447-457.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological methods*, *21*, 475-493.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological methods*, *21*, 458-474.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review, 34,* 135-174.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101,* 958-975.

Flick, C. (2016). Informed consent and the Facebook emotional manipulation study. *Research Ethics, 12,* 14-28.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society, 3,* 1-14

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in cognitive science, 8,* 548-568.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological methods, 21,* 458-474.

Guzzo, R. A., Fink, A. A., King, E., Tonidandel, S., & Landis, R. S. (2015). Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology, 8,* 491-508.

Whelan, T. J., & DuVernet, A. M. (2015). The big duplicity of big data. Industrial and Organizational Psychology, 8(4), 509-515.

Oswald, F. L., & Putka, D. J. (2015). Statistical methods for big data: A scenic tour. In Big Data at Work (pp. 57-77). Routledge.

Braun, M. T., & Kuljanin, G. (2015). Big data and the challenge of construct validity. *Industrial and Organizational Psychology, 8,* 521-527.
Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 165, 234-246.

Sauer, S., Lemke, J., Zinn, W., Buettner, R., & Kohls, N. (2015). Mindful in a random forest: Assessing the validity of mindfulness items using random forests methods. *Personality and Individual Differences, 81,* 117-123.

Ioannidis, J. P. (2013). Informed consent, big data, and the oxymoron of research that is not research. *The American Journal of Bioethics. 13,* 40-42

Strauß, S., & Nentwich, M. (2013). Social network sites, privacy and the blurring boundary between public and private spaces. *Science and Public Policy, 40,* 724-732.

Culpepper, S. A., & Aguinis, H. (2011). R is for revolution: A cutting-edge, free, open source statistical package. Organizational Research Methods, 14, 735-740.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54,* 547-577.

Appendix 1: Project Proposal

Names and G-number:

- A
- B

Research Question:

- Fill in your research question here

Background and Prior Work:

- Fill in your background and prior work here. Be sure to specify which statements are from which references.

Data:

- Explain what the ideal dataset you would want to answer this question. (This should include: What variables? How many observations? Who/what/how would these data be collected? How would these data be stored/organized?)

Team Expectations:

- Team Expectation 1
- Team Expectation 2

Project Timeline Proposal:

Meeting Date	Meeting Time	Completed Before Meeting	Discuss at Meeting

Appendix 2:Final Project Checklist

You can use this checklist to help guide your thinking on the final project. If you check off all the boxes below, you should be in good shape to get a perfect score on your final project.

Abstract:

- Write a clear summary of what you did
- Briefly describe the results of your project

Research Question:

- Include a specific, clear data science question
- Make sure what you're measuring (variables) to answer the question is clear

Background & Prior Work:

- Include a general introduction to your topic
- Include explanation of what work has been done previously
- Include citations or links to previous work

Dataset(s):

- Include an explanation of dataset(s) used (i.e. features/variables included, number of observations, information in dataset)
- Source included (if outside dataset(s) being used)

Data Cleaning & Pre-processing:

- Perform Data Cleaning and explain steps taken OR include an explanation as to why data cleaning was unnecessary (how did you determine your dataset was ready to go?)
- Dataset actually clean and usable after data wrangling steps carried out

Data Visualization:

- Include at least three visualizations
- Clearly label all axes on plots
- Type of all plots appropriate given data displayed
- Interpretation of each visualization included in the text

Data Analysis & Results:

- Exploratory data analysis (EDA) carried out with explanations of what was done and interpretations of output included
- Appropriate analysis performed
- Output of analysis interpreted and interpretation included in notebook

Privacy/Ethics Considerations:

- Discussion of ethical concerns included

Conclusion & Discussion:

- Clear conclusion (answer to the question being asked) and discussion of results
- Limitations of analysis discussed
- Does not ramble on beyond providing necessary information