**Important Dates**

1. November 2nd, 2017 (by 11.59pm) – Project Proposal is due (uploaded to Blackboard by each team member: Course Content > Group Final Project > Project Proposal)
2. November 27th, 2017 – final RMarkdown / Shiny code emailed to course instructor (lewis.tomalin@mountsinai.org)
3. November 30th, 2017 – presentation of the results

*(Choose* ***either*** *statistics project or machine-learning project below)*

**Statistics Project**

**Data Sources**

1. Asthma Call-Back Adult and Child Survey 2006 – 2013 (*ASTHMA*)
   1. Folder: “Data Asthma 2006-2013”
   2. CDC Data: <http://www.cdc.gov/brfss/acbs/index.htm>
   3. Data were collected monthly (by State)
   4. Separate data for Adults and Children
   5. Data saved as .spss datasets
2. PM 2.5 Measures 2003 – 2013 (*PM25*)
   1. Folder: “Data PM2.5 2003-2011”
   2. CDC data: [https://wonder.cdc.gov/wonder/help/PM.html#](https://wonder.cdc.gov/wonder/help/PM.html)
   3. Data were collected monthly (by State)
   4. Data saved as .txt files
3. Ozone Measures 2006 – 2013 (*OZONE*)
   1. Folder: “Data Ozone 2006-2013”
   2. EPA Data: <https://www.epa.gov/outdoor-air-quality-data/download-daily-data>
   3. Data were collected daily (by State)
   4. Data saved as .csv files

**Project Instructions**

Objective: To see how behavior of people with asthma is affected by PM2.5 and Ozone levels.

Each folder has data dictionaries to help you understand the data sets.

Steps:

1. Get familiar with the data:
   1. for *ASTHMA* refer to the “Codebooks Asthma” folder;
   2. for *PM25* – “Outdoor Air Quality - Fine Particulate Matter Doc.pdf” ;
   3. for *OZONE* – “AirData Download Files Documentation.pdf”
2. Design your study:
   1. The outcomes will be 3 variables of your choice in the ASTHMA data set. Choose a two-year time period (not necessarily consecutive) for which the analysis would be based on. Make sure to only include subjects who have asthma.
   2. The explanatory variables will be PM2.5 levels and ozone levels corresponding to your time period.
   3. For each outcome, propose a research question that explores the relationship between the outcome and each of the explanatory variables. State which statistical method you would use to answer the research question. (e.g., t-test, chi-squared test, linear regression). Continuous variables can be turned into categorical variables. Your project questions do not need to be sophisticated.
3. Prepare data for analysis:
4. For each time period, create a “Table 1” for the sample. In other words, provide statistics that summarize the patient characteristics (e.g., age, gender).
5. For each time period, create univariate summary statistics and plots for the 3 asthma outcomes and 2 explanatory variables, PM2.5 and ozone.
6. For each time period, create bivariate summary statistics and plots to show the relationship between the 3 asthma outcomes with each of the 2 explanatory variables, PM2.5 and ozone. In addition, create a plot of the two explanatory variables.
7. For each time period, apply statistical methods to answer your research questions.
8. Descriptively, compare how the two time periods differ in results.
9. Suggestions on how to divide responsibilities among the team members:
   1. Create project write-up
   2. Import and clean the *ASTHMA* datasets
   3. Import and clean the *OZONE* datasets
   4. Import and clean the *PM25* datasets
   5. Merge data sets
   6. Create univariate and bivariate basic descriptive statistics
   7. Conduct formal statistical tests to answer research questions
   8. Create RMarkdown or Shiny application
10. Final project RMarkdown / Shiny code and output will have to be emailed to the course instructor (Maria Suprun) by November 28th, 2016. You should include percent effort contributed by each team member.
11. You will present your project results on the last day of class (December 1, 2016). You should use RMarkdown or Shiny to present your results.

**Machine-Learning Project**

**Data Sources**

1. **Treatment response of psoriasis patients to ETANERCEPT – 2009 (GSE11903)**
   1. Originally cited: “Zeba et al, 2009, PMID: 19895991”
   2. GEO Summary: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11903>.
   3. Time Points: Time0, Wk1, Wk2, Wk4.
   4. Treatment response is in pData (Responder=R, Non-responder=NR)
   5. Lesional (LS) and non-lesional (NL) skin samples collected.
2. **Response of psoriasis patients to ADALIMUMAB and METHOTREXATE – (GSE85034)** 
   1. Originally cited: “Goldminz et al, 2015 & 2016, PMID: 25946554; PMID: 27005302”
   2. GEO Summary: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85034>
   3. Time Points: Time0, Wk1, Wk2, Wk4, Wk16
   4. PASI score recorded for each sample (if patient PASI score improves 75% by wk16, deem patient a responder)
   5. Lesional (LS) and non-lesional (NL) skin samples collected.
3. **Treatment response to TOFACITINIB and PLACEBO – 2016 (GSE69967)**
   1. Originally cited: “Krueger et al, 2016, PMID: 27059729”
   2. GEO Summary: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69967>
   3. Time Points: Day 0, Day 1, Day 3, Week 1, Week 2, Week 4, Week 12.
   4. Treatment response: Wk12 PASI improvement included in pData (PASI improvement > 75% deem as responder)

**Project Instructions**

Objective: Develop three machine-learning pipelines to predict treatment response from gene-expression data.

Data sets are summarized on GEO website, also read original citations.

Steps:

1. Get familiar with the data:
   1. for *ETANERCEPT* - “Zeba et al, 2009, PMID: 19895991”
   2. for *ADALIMUMAB* - “Goldminz et al, 2015, PMID: 25946554”
   3. for *METHOTREXATE* -: “Goldminz et al, 2015/2016, PMID: 25946554/27005302”
   4. for *TOFACITINIB & PLACEBO* - “Krueger et al, 2016, PMID: 27059729”
2. Design your machine-learning pipeline:
   1. Your outcome variable (Y) will be treatment response (responder or non-responder).
   2. Your features (X) with be a matrix/data.frame of gene-expression for each patient.
   3. Use as many (or as few) samples as you would like (eg time-points, genes, LS/NL, patients, drugs etc).
   4. Decide which ML algorithms you want to use for training (eg: glmnet, nnet, pls (see Caret for list of methods)).
   5. You can try to improve your predictive performance with resampling, bootstrapping and bagging or any other methods you know.
   6. You may split your data into training and test sets (the proportion is up to you).
   7. Measure performance of each method against a small validation set (~10% of patients).
   8. Use the same validation set for each method.
   9. Validation patients should **not** be used in training, resampling bagging etc.
   10. Describe your strategies both graphically and verbally.
3. Prepare data for analysis:
4. For each treatment summarize the available pData, number of patients, time points, number of responders and non-responders etc.
5. Filter your data anyway you wish (eg remove low expression genes).
6. Transform your expression data so that each observation corresponds to one patient rather than one sample.
7. Summarize overall gene-expression for different treatments, time points etc.
8. Apply your three strategies.
9. Provide multiple performance metrics, including plots.
10. Descriptively and graphically compare the performance of each strategy.
11. Suggestions on how to divide responsibilities among the team members:
    1. Create project write-up
    2. Import, clean and transform datasets
    3. Summarize data sets
    4. Write/run script for strategy #1
    5. Write/run script for strategy #2
    6. Write/run script for strategy #3
    7. Summarize and compare predictive performance
    8. Create RMarkdown or Shiny application
12. Final project RMarkdown / Shiny code and output will have to be emailed to the course instructor (Lewis Tomalin) by November 27th, 2017. You should include percent effort contributed by each team member.
13. You will present your project results on the last day of class (December 30, 2017). You should use RMarkdown or Shiny to present your results.

**Group Project Guidelines**

Five to seven students will be assigned in a group. Groups will have to prepare a reproducible report/ presentation / web application for the project using either RMarkdown or Shiny.

Groups should submit a Group Project Proposal by November 2nd (**Lecture 9)** by uploading it to Blackboard (each group member will have to upload it separately). Proposal should be no more than 1 page long and include the following:

1. Project title
2. Group members
3. Objectives (i.e., research questions)
4. Brief description of your approach

Project presentations will be on the last day of class (**Lecture 12**). Each group will have 10 minutes to present their application, followed by a 3-minute Q&A.

This project is an opportunity for you to demonstrate everything you have learned in this module. Higher marks will be awarded to groups who demonstrate a diverse array of programming techniques.