# Course Outline

## Day 1

1. **Introduction to Machine Learning and Data Science (1 hour)**
    a. The Big Data Revolution and the rise of data science
    b. New sources of data and the value of interconnected datasets
    c. Where the magic happens: the importance of analytics and machine learning in order to deliver concise actionable items from big and complex data
    d. What is statistical learning and how it relates to econometrics and statistics?
        i. The main analytic paradigms: exploration, prediction, explanation
        ii. The importance of social science: addressing heterogeneity and selection bias
        iii. Machine learning and ethical dilemmas
    e. Examples of machine learning:
        i. Creating new datasets (mining text and images)
        ii. Pattern mining and generating hypotheses (dimensionality reduction)
        iii. Using machine learning to improve existing econometric strategies (causal inference)
        iv. Using machine learning to develop new tools (neural networks and deep learning)
2. ***Overview of computational tools (0.5 hours)***
    a. R and Python; Jupyter Notebooks
    b. First hands-on experience with sample data for this course:
        i. Administrative data: property records
        ii. Transaction data: online shopping data
        iii. Program data: behavioral marketing data
3. **Core concepts in statistical learning (1 hour)**
    a. Supervised and unsupervised learning
    b. Review: linear regression (estimation, prediction, model specification)
    c. Trade-offs: prediction accuracy vs interpretability
    d. The bias-variance trade-off
    e. Addressing the overfitting problem:
        i. Validation sets
        ii. K-fold cross-validation
        iii. Bootstrap and boosting
4. ***Demonstration in R: Linear regression and data driven inference in R (0.5 hours)***
    a. Multiple regression analysis
    b. Choosing the functional form of a model using k-fold cross-validation in polynomial regression
5. **Regression and prediction (2 hours)**
    a. Regression in high dimensions
    b. Model selection
    c. Regularization
        i. Penalized objective functions
        ii. Ridge regression, LASSO, Elastic net
    d. Polynomial regression and regression splines
6. ***Demonstration in R: Regression analysis (0.5 hours)***

    a. Building a hedonic model to predict house prices

7. **Classification (1 hour)**
    a. Discrete and binary data
    b. Logistic models: Binary responses and multiple responses
    c. Discriminant analysis
    d. Evaluating performance: confusion matrix
8. ***Demonstration in R: Classification (0.5 hours)***
    a. Predicting customers' purchase decisions in an online store

# Day 2

1. **Unsupervised learning (1 hour)**
    a. Principal Component Analysis and Factor Models
    b. Clustering methods: k-Means clustering, hierarchical clustering
    c. Examples:
        i. Customer segmentation and latent consumer preferences for CPG brands
        ii. Constructing a control group for policy analysis; clustering road segment data
2. ***Demonstration in R: Unsupervised learning (0.5 hours)***
    *a.* PCA and k-means clustering on scanner data for online grocery purchases
3. **Introduction to Causal Inference (1.5 hours)**
    a. Rubin Causal Model
    b. Experiments and Randomization
        i. Example: Evaluating online advertising in large scale field experiments
    c. Natural experiments and Difference-in-Differences
        i. Example: Evaluating behavioral marketing with random adoption
    d. Propensity Score Models
        i. Example: Measuring customer response to opt-in programs
    e. Instrumental Variables
4. **From Prediction to Causal Inference (1 hour)**
    a. Why machine learning leads to biased structural estimated?
    b. Estimating treatment effects in high-dimensional models: what to do when you have too many control variables?
    c. Post-selection inference in high-dimensional causal models
    d. Double-machine learning
5. ***Demonstration in R: Double Machine Learning (0.75 hours)***
    a. Evaluating customer response to a behavioral marketing program
6. **Personalization and Heterogeneous Treatment Effects (1 hour)**
    a. Quantile Regression Analysis
    b. Decision Trees and random forests
    c. Using random forests to estimate heterogeneous treatment effects
7. ***Demonstration in R: Random forests and causal inference (0.75 hours)***
    a. Evaluating heterogeneous customer response to a behavioral marketing program
8. **Next steps (0.5 hours)**
    a. Working with data scientists and computer scientists: effective team building
    b. Machine learning at scale
    c. Recent advances: neural networks and deep learning, economics and artificial intelligence